(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 558 124**

(51) Int. CI.:

C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **20.12.2006 E 14166884 (8)**

(97) Fecha y número de publicación de la concesión europea: **16.12.2015 EP 2789696**

(54) Título: **Procedimiento para detección de polimorfismos basada en AFLP de alto rendimiento**

(30) Prioridad:

22.12.2005 US 752590 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
02.02.2016

(73) Titular/es:

KEYGENE N.V. (100.0%)
P.O. Box 216
6700 AE Wageningen, NL

(72) Inventor/es:

VAN EIJK, MICHAEL JOSEPHUS THERESIA;
SØRENSEN, ANKER PREBEN y
VAN SCHRIEK, MARCO GERARDUS MARIA

(74) Agente/Representante:

PONTI SALES, Adelaida

DESCRIPCIÓN

Procedimiento para detección de polimorfismos basada en AFLP de alto rendimiento

Campo técnico

[0001] La presente invención se refiere a los campos de la biología molecular y la genética. La invención se refiere a un descubrimiento rápido, detección y genotipificación a gran escala de polimorfismos en una muestra de ácido nucleico o entre muestras. Los polimorfismos identificados pueden usarse como marcadores genéticos.

Antecedentes de la invención

[0002] La exploración del ADN genómico ha sido pretendida desde hace tiempo por la comunidad científica, en particular la médica. El ADN genómico contiene la clave de la identificación, el diagnóstico y el tratamiento de enfermedades como el cáncer y la enfermedad de Alzheimer. Además de la identificación y tratamiento de enfermedades, la exploración del ADN genómico puede proporcionar ventajas importantes en los esfuerzos de mejora de animales y plantas, que pueden proporcionar respuestas a problemas alimentarios y de nutrición en el mundo.

[0003] Se sabe que muchas enfermedades están asociadas con componentes genéticos específicos, en particular con polimorfismos en genes específicos. La identificación de polimorfismos en grandes muestras como los genomas es en la actualidad una tarea laboriosa y larga. Sin embargo, dicha identificación tiene gran valor para áreas como la investigación biomédica, el desarrollo de productos farmacéuticos, la tipificación de tejidos, la genotipificación y los estudios de población.

[0004] Los marcadores, en particular marcadores genéticos, se han usado desde hace mucho tiempo como un procedimiento de tipificación genética, es decir, para relacionar un rasgo fenotípico con la presencia, ausencia o cantidad de una parte de ADN (gen) en particular. El documento WO-2005/003.375 desvela un procedimiento para tipificación genética de un único genoma (haplotipificación) que comprende (1) preparación de muestras de ácido nucleico, (2) amplificación de ácidos nucleicos, y (3) secuenciación de ADN en un tubo de reacción. Otro procedimiento es AFLP, que es una de las tecnologías más versátiles de tipificación genética ya desde hace muchos años y que puede aplicarse ampliamente a cualquier organismo (para revisiones véase Savelkoul y col. J. Clin. Microbiol, 1999, 37(10), 3083-3091; Bensch y col. Molecular Ecology, 2005, 14, 2899-2914).

[0005] La tecnología AFLP (Zabeau & Vos, 1993; Vos *y col.,* 1995) ha encontrado un uso extenso en mejora vegetal y otros campos desde su invención a principios de los años noventa. Ello se debe a varias características de la AFLP, entre las cuales la más importante es que no se necesita una información previa de las secuencias para generar un gran número de marcadores genéticos de una forma reproducible. Además, el principio de la amplificación selectiva, una piedra angular de la AFLP, asegura que el número de fragmentos amplificados puede ponerse en línea con la resolución del sistema de detección, con independencia del tamaño o el origen del genoma.

[0006] La detección de fragmentos de AFLP se realiza habitualmente mediante electroforesis en placa de gel (Vos *y col.,* 1995) o electroforesis capilar (van der Meulen *y col.,* 2002). La mayoría de los marcadores de AFLP valorados de esta forma representan polimorfismos (de un solo nucleótido) que se producen en los sitios de reconocimiento de enzimas de restricción usados para la preparación de la plantilla de AFLP o en sus nucleótidos de flanqueo cubiertos por cebadores de AFLP selectivos. El resto de los marcadores de AFLP son polimorfismos de inserción/deleción que aparecen en las secuencias internas de los fragmentos de restricción y en una fracción muy pequeña en sustituciones de nucleótidos individuales que tienen lugar en fragmentos de restricción pequeños (< aproximadamente 100 pb), que para estos fragmentos provocan variaciones de movilidad reproducibles entre los dos alelos; estos marcadores de AFLP pueden valorarse de forma codominante sin tener que depender de las intensidades de banda.

[0007] En una huella genética típica de AFLP, los marcadores de AFLP constituyen por tanto la minoría de los fragmentos amplificados (menos del 50% y a menudo menos del 20%), mientras que el resto se refieren habitualmente como fragmentos de AFLP constantes. No obstante, los últimos son útiles en el procedimiento de valoración en gel ya que sirven como puntos de sujeción para calcular las movilidades de los fragmentos de los marcadores de AFLP y ayudan a cuantificar los marcadores para valoración codominante. La valoración codominante (valoración para homocigosidad o heterocigosidad) de marcadores de AFLP está limitada en la actualidad al contexto de la obtención de huellas moleculares para segregación de poblaciones. En un panel de

líneas no relacionadas sólo es posible una valoración dominante.

[0008] Aunque el rendimiento de AFLP es muy alto debido a los altos niveles de multiplexado en las etapas de amplificación y detección, la etapa limitativa de la velocidad es el poder de resolución de la electroforesis. La electroforesis permite una identificación única de la mayoría de los fragmentos amplificados basándose en la combinación de combinaciones de enzimas de restricción (EC), combinaciones de cebadores (PC) y movilidad, aunque idealmente, el sistema de detección debe ser capaz de determinar la secuencia completa de los fragmentos amplificados para capturar todos los polimorfismos.

[0009] La detección por secuenciación en lugar de la determinación de la movilidad aumentará el rendimiento porque:

1) los polimorfismos situados en las secuencias internas serán detectados en la mayoría (o la totalidad) de los fragmentos amplificados; así se incrementará considerablemente el número de marcadores por PC;

2) no existe pérdida de marcadores de AFLP debido a la comigración de marcadores de AFLP y las bandas constantes;

3) la valoración codominante no depende de la cuantificación de las intensidades de bandas y es independiente del parentesco de las personas para las que se ha tomado la huella genética.

[0010] Hasta ahora, la detección de marcadores de AFLP/secuencias por secuenciación no ha sido viable económicamente debido, entre otras limitaciones, a limitaciones de costes de la tecnología de secuenciación didesoxi de Sanger y otras tecnologías de secuenciación convencionales.

[0011] En consecuencia, uno de los objetivos de la presente invención es proporcionar procedimientos económicamente viables para la detección de marcadores de AFLP u otros marcadores genéticos tales como marcadores de SNP basados en la secuenciación.

[0012] Un problema importante adicional asociado con la detección de una colección de fragmentos que contienen AFLP o SNP a través de la secuenciación para fines de genotipificación (es decir, de diagnóstico) es la variación del muestreo. Específicamente, esto significa que cuando se analiza una colección de fragmentos y no se observan fragmentos especiales, es preciso cerciorarse de que no se debe al hecho de que los fragmentos implicados no fueron muestreados en la etapa de detección, aunque estén presentes en la mezcla de fragmentos, ya que ello conduciría a una consideración como falsos negativos del marcador. Esta limitación no se aplica a la detección por electroforesis ya que se dispone de la información de posición en el gel.

[0013] En consecuencia, uno de los objetivos adicionales de la presente invención es proporcionar un procedimiento que resuelva el problema de la variación de muestras o al menos reduzca los errores por variación de muestras a un mínimo aceptable.

Resumen de la invención

[0014] Los autores de la presente invención han encontrado que la secuenciación se encuentra dentro del alcance de la detección de marcadores de AFLP y SNP con el uso de AFLP en ciertos procedimientos adaptados para secuenciación de alto rendimiento. La invención proporciona así un procedimiento o estrategia que combina el poder y la aplicabilidad genérica de la AFLP con determinadas tecnologías de secuenciación de alto rendimiento para establecer un sistema de valoración de polimorfismos aplicable genéticamente. En esta estrategia, se aborda también la cuestión de la variación del muestreo para garantizar la genotipificación con alta precisión y elevar al máximo las posibilidades de obtener conjuntos de datos con números mínimos de genotipos ausentes.

Definiciones

[0015] En la siguiente descripción y en los ejemplos propuestos se usan distintos términos. Con el fin de proporcionar una comprensión clara y coherente de la memoria descriptiva y las reivindicaciones, incluido el ámbito en el que se incluyen dichos términos, se proporcionan las siguientes definiciones. A menos que se defina de otro modo en la presente memoria descriptiva, todos los términos técnicos y científicos usados poseen el mismo significado que entienden normalmente los expertos en la materia a los que se dirige la presente invención.

[0016] Polimorfismo: polimorfismo se refiere a la presencia de dos o más variantes de una secuencia de nucleótidos en una población. Un polimorfismo puede comprender uno o más cambios de base, una inserción, una repetición o una deleción. Un polimorfismo incluye por ejemplo una repetición de secuencia simple (SSR) y un polimorfismo de nucleótido simple (SNP), que es una variación, que tiene lugar cuando se altera un solo nucleótido: adenina (A), timina (T), citosina (C) o guanina (G). Una variación debe producirse en general en al menos el 1% de la población para que se considere SNP. Los SNP conforman por ejemplo el 90% de todas las variaciones genéticas humanas, y se producen en cada 100 a 300 bases a lo largo del genoma humano. Dos de cada tres SNP sustituyen la citosina (C) por timina (T). Las variaciones en las secuencias de ADN de, por ejemplo, seres humanos o plantas pueden influir en el modo de manejar las enfermedades, las bacterias, los virus, los productos químicos, los fármacos, etc.

[0017] Ácido nucleico: un ácido nucleico según la presente invención puede incluir cualquier polímero u oligómero de bases pirimidina o purina, preferentemente citosina, timina, y uracilo, y adenina y guanina, respectivamente (Véase Albert L. Lehninger, Principles of Biochemistry, en 793-800 (Worth Pub. 1982)). La presente invención contempla cualquier componente de ácido nucleico de desoxirribonucleótido, ribonucleótido o péptido, y cualquier variante química de los mismos, tal como formas metiladas, hidroximetiladas o glucosiladas de estas bases, y similares. Los polímeros u oligómeros pueden ser de composición heterogénea u homogénea, y pueden aislarse a partir de fuentes de ocurrencia natural o pueden producirse de forma artificial o sintética. Además, los ácidos nucleicos pueden ser ADN o ARN, o una mezcla de los mismos, y pueden existir de forma permanente o en transición en forma monocatenaria o bicatenaria, lo que incluye homodúplex, heterodúplex y estados híbridos.

[0018] Reducción de la complejidad: el término reducción de la complejidad se usa para referirse a un procedimiento en el que la complejidad de una muestra de ácido nucleico, como un ADN genómico, se reduce por la generación de un subconjunto de la muestra. Este subconjunto puede ser representativo de la muestra total (es decir, complejo) y preferentemente es un subconjunto reproducible. Reproducible significa en este contexto que cuando la misma muestra se reduce en complejidad usando el mismo procedimiento, se obtiene el mismo subconjunto, o al menos comparable. El procedimiento usado para la reducción de la complejidad puede ser cualquier procedimiento para la reducción de la complejidad conocido en la técnica. Un ejemplo preferido de un procedimiento para reducción de la complejidad incluye por ejemplo AFLP® (Keygene N.V., Países Bajos; véanse por ejemplo los documentos EP-0.534.858, US-60.459.94), los procedimientos descritos por Dong (véanse por ejemplo los documentos WO-03/012.118, WO-00/24.939), unión indexada (Unrau y col., ver más abajo), PCR de ligador (documento WO-90/008.821) y SALSA-PCR (documento WO-00/23.620) Schouten y col.) etc. Los procedimientos de reducción de la complejidad usados en la presente invención tienen en común que son reproducibles. Reproducibles en el sentido de que cuando se reduce en complejidad la misma muestra de la misma forma, se obtiene el mismo subconjunto de la muestra, a diferencia de una reducción de la complejidad más aleatoria tal como la microdisección o el uso de ARNm (ADNc) que representa una parte del genoma transcrito en un tejido seleccionado y para su reproducibilidad depende de la selección del tejido, el tiempo de aislamiento, etc.

[0019] AFLP: AFLP se refiere a un procedimiento para amplificación selectiva de ADN basada en la digestión de un ácido nucleico con una o más endonucleasas de restricción para producir fragmentos de restricción, que ligan los adaptadores a los fragmentos de restricción y amplifican los fragmentos de restricción ligados a adaptadores con al menos un cebador que es (parte) complementaria al adaptador, **(**parte) complementaria a los restos de la endonucleasa de restricción, y que contiene además al menos un nucleótido seleccionado aleatoriamente entre A,C, T o G (o U según cuál sea el caso). AFLP no necesita ninguna información de secuencia anterior y puede realizarse en cualquier ADN de inicio. En general, AFLP comprende las etapas consistentes en:

(a) digestión de un ácido nucleico, en particular un ADN o ADNc, con una o más endonucleasas de restricción específicas, para fragmentar el ADN en una serie correspondiente de fragmentos de restricción;

(b) unión de los fragmentos de restricción así obtenidos con un adaptador de oligonucleótido sintético bicatenario, uno de cuyos extremos es compatible con uno o los dos extremos de los fragmentos de restricción, para producir en consecuencia fragmentos de restricción ligados al adaptador, preferentemente etiquetados, del ADN de inicio;

(c) puesta en contacto de los fragmentos de restricción ligados al adaptador, preferentemente etiquetados en condiciones de hibridación con uno o más cebadores de oligonucleótidos que contienen nucleótidos selectivos en su extremo 3';

(d) amplificación del fragmento de restricción ligado al adaptador, preferentemente etiquetado hibridado con los cebadores por PCR o una técnica similar de manera que se induzca una mayor elongación de los cebadores

hibridados a lo largo de los fragmentos de restricción del ADN de inicio con el que se hibridan los cebadores; y

(e) detección, identificación o recuperación del fragmento de ADN amplificado o alargado así obtenido.

[0020] La AFLP proporciona así un subconjunto reproducible de fragmentos ligados a adaptador. La AFLP se describe en los documentos EP-534.858, US-6.045.994 y en Vos *y col.* Se hace referencia a estas publicaciones para más detalles sobre la AFLP. La AFLP se usa habitualmente como una técnica de reducción de la complejidad y una tecnología de obtención de huellas genéticas. Dentro del contexto del uso de la AFLP como tecnología de obtención de huellas genéticas, se ha desarrollado el concepto de marcador de AFLP.

[0021] Marcador de AFLP: Un marcador de AFLP es un fragmento de restricción ligado a adaptador amplificado que es diferente entre dos muestras que han sido amplificadas usando AFLP (con huella genética), usando el mismo conjunto de cebadores. De este modo, la presencia o ausencia de este fragmento de restricción ligado a adaptador amplificado puede usarse como un marcador que está unido a un rasgo o fenotipo. En tecnología de gel convencional, un marcador de AFLP se muestra como una banda en el gel situada a una movilidad determinada. Otras técnicas de electroforesis tales como electroforesis capilar pueden no referirse a ésta como banda, si bien el concepto sigue siendo el mismo, es decir, un ácido nucleico con cierta longitud y movilidad. La ausencia o presencia de la banda puede ser indicativa de (o estar asociada con) la presencia o ausencia del fenotipo. Los marcadores de AFLP implican normalmente SNP en el sitio de restricción de la endonucleasa o los nucleótidos selectivos. En ocasiones, los marcadores de AFLP pueden implicar indeles (inserciones-deleciones) en el fragmento de restricción.

[0022] Marcador de SNP: un marcador de SNP es un marcador que se basa en un polimorfismo de nucleótido simple identificado en una cierta posición. Los marcadores de SNP pueden estar situados en posiciones idénticas a los marcadores de AFLP, si bien los marcadores de SNP también pueden estar situados en el propio fragmento de restricción. De este modo los marcadores de SNP genéricos comprenden así la especie los marcadores de AFLP.

[0023] Banda constante: una banda constante en la tecnología de AFLP es un fragmento de restricción ligado a adaptador amplificado que es relativamente invariable entre muestras. Así, una banda constante en la tecnología de AFLP se mostrará, en un intervalo de muestras, aproximadamente en la misma posición en el gel, es decir, tiene la misma longitud/movilidad. En AFLP convencional se usa normalmente para fijar las vías correspondientes a las muestras en un gel o electroferogramas de múltiples muestras de AFLP detectadas por electroforesis capilar. Normalmente, una banda constante es menos informativa que un marcador de AFLP. No obstante, como los marcadores de AFLP implican normalmente SNP en los nucleótidos selectivos o en el sitio de restricción, las bandas constantes pueden comprender SNP en los propios fragmentos de restricción, lo que hace de las bandas constantes una interesante fuente alternativa de información genética que es complementaria a los marcadores de AFLP.

[0024] Base selectiva: situada en el extremo 3' del cebador que contiene una parte que es complementaria al adaptador y una parte que es complementaria al resto del sitio de restricción, la base selectiva se selecciona aleatoriamente entre A, C, T o G. Al extender un cebador con una base selectiva, la posterior amplificación producirá sólo un subconjunto reproducible de los fragmentos de restricción ligados a adaptador, es decir, sólo los fragmentos que pueden ser amplificados usando el cebador que contiene la base selectiva. Los nucleótidos selectivos pueden añadirse en el extremo 3' del cebador en un número comprendido entre 1 y 10. Normalmente bastará entre 1 y 4. Los dos cebadores pueden contener un número variable de bases selectivas. Con cada base selectiva añadida, el subconjunto reduce la cantidad de fragmentos de restricción ligados a un adaptador amplificados en el subconjunto en un factor de aproximadamente 4. Normalmente, el número de bases selectivas usadas en AFLP está indicado por +N+M, en el que un cebador lleva N nucleótidos selectivos y los otros cebadores llevan M nucleótidos selectivos. Así, Eco/Mse +1/+2 AFLP es la abreviatura de la digestión del ADN de inicio con EcoRI y MseI, ligación de adaptadores apropiados y amplificación con un cebador dirigido a la posición restringida EcoRI que lleva una base selectiva y el otro cebador dirigido al sitio restringido MseI que lleva 2 nucleótidos selectivos.

[0025] Agrupación: con el término "agrupación" se indica la comparación de dos o más secuencias de nucleótidos basándose en la presencia de longitudes cortas o largas de nucleótidos idénticos o similares. En la técnica se conocen varios procedimientos para alineación de secuencias de nucleótidos, como se explicará más adelante. A veces los términos "ensamblaje" o "alineación" se usan como sinónimos.

[0026] Etiqueta: una breve secuencia que puede añadirse a un cebador o incluirse en su secuencia o usarse de otro modo como etiqueta para proporcionar un identificador único. Dicho identificador de secuencia puede ser

una base secuencia única de longitud variable pero definida usada en exclusiva para identificar una muestra específica de ácido nucleico. Por ejemplo las etiquetas de 4 pb permiten 4(exp4) = 256 etiquetas diferentes. Algunos ejemplos típicos son las secuencias ZIP, conocidas en la técnica como etiquetas usadas habitualmente para la detección única por hibridación (Iannone y col. Cytometry 39:131-140, 2000). Usando dicha etiqueta, el origen de una muestra de PCR puede determinarse tras un procesamiento adicional. En el caso de combinación de productos procesados que proceden de diferentes muestras de ácido nucleico, las diferentes muestras de ácido nucleico se identifican en general usando diferentes etiquetas. En el caso de la presente invención, la adición de una única etiqueta de secuencia sirve para identificar las coordenadas de la planta individual en el conjunto de productos de amplificación de secuencias. Pueden usarse múltiples etiquetas.

[0027] Etiquetado: el término etiquetado se refiere a la adición de una etiqueta a una muestra de ácido nucleico con el fin de poder distinguirla de una segunda muestra de ácido nucleico o una muestra adicional. El etiquetado puede realizarse por ejemplo por la adición de un identificador de secuencia durante la reducción de la complejidad o por cualquier otro medio conocido en la técnica. Dicho identificador de secuencia puede ser por ejemplo una secuencia de base única de longitud variable pero definida usada únicamente para identificar una muestra de ácido nucleico específica. Los ejemplos típicos de los mismos son por ejemplo secuencias ZIP. Usando dicha etiqueta, el origen de una muestra puede determinarse mediante procesamiento adicional. En caso de combinación de productos procesados de diferentes muestras de ácido nucleico, las diferentes muestras de ácido nucleico deben identificarse usando diferentes etiquetas.

[0028] Biblioteca etiquetada: el término biblioteca etiquetada se refiere a una biblioteca de ácidos nucleicos etiquetados.

[0029] Secuenciación: el término secuenciación se refiere a la determinación del orden de nucleótidos (secuencias de bases) en una muestra de ácidos nucleicos, por ejemplo ADN o ARN.

[0030] Cribado de alto rendimiento: el cribado de alto rendimiento, a menudo abreviado como HTS *(high-throughput screening)*, es un procedimiento para experimentación científica especialmente relevante para los campos de la biología y la química. A través de una combinación de robótica moderna y otro hardware de laboratorio especializado, permite al investigador cribar eficazmente grandes cantidades de muestras simultáneamente.

[0031] Endonucleasa de restricción: una endonucleasa de restricción o enzima de restricción es una enzima que reconoce una secuencia de nucleótidos específica (sitio diana) en una molécula de ADN bicatenaria, y escindirá las dos cadenas de la molécula de ADN en cada sitio diana.

[0032] Fragmentos de restricción: las moléculas de ADN producidas por digestión con una endonucleasa de restricción se refieren como fragmentos de restricción. Cualquier genoma (o ácido nucleico, con independencia de su origen) dado será digerido por una endonucleasa de restricción particular en un conjunto discreto de fragmentos de restricción. Los fragmentos de ADN que proceden de la escisión de la endonucleasa de restricción pueden usarse adicionalmente en diversas técnicas y, por ejemplo, pueden ser detectados por electroforesis en gel.

[0033] Electroforesis en gel: con el fin de detectar fragmentos de restricción, puede necesitarse un procedimiento analítico para fraccionar moléculas de ADN bicatenario basándose en el tamaño. La técnica usada más comúnmente para conseguir dicho fraccionamiento es electroforesis (capilar) en gel. La velocidad a la que se mueven los fragmentos de ADN en dichos geles depende de su peso molecular; así, las distancias recorridas disminuyen cuando aumentan las longitudes de los fragmentos. Los fragmentos de ADN fraccionados por electroforesis en gel pueden visualizarse directamente por un procedimiento de tinción, por ejemplo, tinción con plata o tinción que usa bromuro de etidio, si el número de fragmentos incluidos en el patrón es suficientemente pequeño. Alternativamente el tratamiento adicional de fragmentos de ADN puede incluir etiquetas detectables en los fragmentos, tales como fluoróforos o etiquetas radiactivas.

[0034] Ligación: la reacción enzimática catalizada por una enzima ligasa en la que dos moléculas de ADN bicatenario se unen entre sí por enlaces covalentes se refiere como ligación. En general, las dos cadenas de ADN están unidas entre sí por enlaces covalentes, aunque también es posible evitar la ligación de una de las dos cadenas a través de modificación química o enzimática de uno de los extremos de las cadenas. En ese caso la unión covalente tendrá lugar en sólo una de las dos cadenas de ADN.

[0035] Oligonucleótido sintético: las moléculas de ADN monocatenario que tienen preferentemente de aproximadamente 10 a aproximadamente 50 bases, que pueden sintetizarse químicamente, se refieren como

oligonucleótidos sintéticos. En general, estas moléculas de ADN sintéticas están diseñadas de manera que tienen una secuencia de nucleótidos única o deseada, aunque es posible sintetizar familias de moléculas que tienen secuencias relacionadas y que tienen diferentes composiciones de nucleótidos en posiciones específicas dentro de la secuencia de nucleótidos. El término oligonucleótido sintético se usará para referirse a moléculas de ADN que tienen una secuencia de nucleótidos designada o deseada.

[0036] Adaptadores: moléculas cortas de ADN bicatenario con un número limitado de pares de bases, por ejemplo de aproximadamente 10 a aproximadamente 30 pares de bases de longitud, que están diseñadas de manera que pueden ligarse a los extremos de fragmentos de restricción. Los adaptadores están compuestos en general por dos oligonucleótidos sintéticos que tienen secuencias de nucleótidos que son parcialmente complementarias entre sí. Cuando se mezclan los dos oligonucleótidos sintéticos en solución en condiciones apropiadas, se aparean entre sí para formar una estructura bicatenaria. Después del apareamiento, un extremo de la molécula del adaptador está diseñado de manera que es compatible con el extremo de un fragmento de restricción y puede estar ligado al mismo; el otro extremo del adaptador puede diseñarse de manera que no puede ligarse, si bien esto no tiene por qué suceder (adaptadores ligados dobles).

[0037] Fragmentos de restricción ligados a adaptador: fragmentos de restricción que han sido rematados por adaptadores.

[0038] Cebadores: en general, el término cebadores se refiere a cadenas de ADN que pueden cebar la síntesis de ADN. La ADN polimerasa no puede sintetizar ADN *de novo* sin cebadores: sólo puede extender una cadena de ADN existente en una reacción en la que la cadena complementaria se usa como plantilla para dirigir el orden de los nucleótidos que se van a ensamblar. Se hará referencia a las moléculas de oligonucleótidos sintéticos que se usan en una reacción en cadena de la polimerasa (PCR) como cebadores.

[0039] Amplificación de ADN: el término amplificación de ADN se usará normalmente para referirse a la síntesis *in vitro* de moléculas de ADN bicatenario usando PCR. Debe observarse que existen otros procedimientos de amplificación y pueden usarse en la presente invención sin alejarse de su espíritu.

[0040] Hibridación selectiva: se refiere a hibridación, en estrictas condiciones de hibridación, de una secuencia de ácidos nucleicos a una secuencia diana de un ácido nucleico especificado con un grado detectablemente superior (por ejemplo, al menos 2 veces con respecto al fondo) al de su hibridación con secuencias no diana de ácidos nucleicos y con la exclusión sustancial de ácidos nucleicos no diana. Los términos "condiciones de fidelidad" o "condiciones de fidelidad de hibridación" incluyen referencia a condiciones según las cuales una sonda se hibridará con su secuencia diana, en un grado detectablemente superior que otras secuencias (por ejemplo, al menos 2 veces con respecto al fondo). Las condiciones de fidelidad dependen de la secuencia y serán diferentes en distintas circunstancias. Mediante el control de la fidelidad de la hibridación y/o las condiciones de lavado, las secuencias diana pueden identificarse como complementarias al 100% con la sonda (sondeo homólogo). Alternativamente, las condiciones de fidelidad pueden ajustarse para permitir cierto malapareamiento en las secuencias de manera que se detecten menores grados de semejanza (sondeo heterólogo). En general, una sonda es inferior a aproximadamente 100 nucleótidos de longitud, opcionalmente con no más de 50, o 25 nucleótidos de longitud. Normalmente, las condiciones de fidelidad serán aquellas en las que la concentración de sales es menor que aproximadamente 1,5 M de iones Na, normalmente una concentración de aproximadamente 0,01 a 1,0 M de iones Na (u otras sales) a pH de 7,0 a 8,3 y la temperatura es de al menos aproximadamente 30°C para sondas cortas (por ejemplo, de 10 a 50 nucleótidos) y al menos aproximadamente 60°C para sondas largas (por ejemplo, superior a 50 nucleótidos). Las condiciones de fidelidad pueden alcanzarse también con la adición de agentes de desestabilización como la formamida. Las condiciones de ejemplo de baja fidelidad incluyen hibridación con una solución tampón del 30 al 35% de formamida, NaCl 1 M, SDS (dodecilsulfato de sodio) al 1% a 37°C, y un lavado en 1* a 2*SSC (20*SSC=3,0 M NaCl/0,3 M citrato de trisodio) a entre 50 y 55°C. Entre los ejemplos de condiciones de fidelidad moderadas se incluyen hibridación entre el 40 y el 45% de formamida, NaCl 1 M, SDS al 1% a 37°C y un lavado en 0,5* a 1*SSC a entre 55 y 60°C. Entre las condiciones de ejemplo de alta fidelidad se incluyen hibridación en formamida al 50%, NaCl 1 M, SDS al 1% a 37°C y un lavado en 0,1*SSC a entre 60 y 65°C. La especificidad es normalmente función de lavados post-hibridación, siendo los factores críticos la resistencia iónica y la temperatura de la solución final de lavado. Para híbridos ADN-ADN, la Tm puede aproximarse a partir de la ecuación de Meinkoth y Wahl, Anal. Biochem., 138:267-284 (1984): Tm = 81,5°C + 16,6 (log M) + 0,41 (% GC)-0,61 (% form) -500/L; en la que M es la molaridad de cationes monovalentes, % GC es el porcentaje de nucleótidos de guanosina y citosina en el ADN, % form es el porcentaje de formamida en la solución de hibridación, y L es la longitud del híbrido en pares de bases. La Tm es la temperatura (con resistencia iónica y pH definidos) a la que el 50% de una secuencia diana complementaria se hibrida con una sonda perfectamente correspondiente. Tm se reduce aproximadamente en 1°C

por cada 1% de malapareamiento; así, Tm, la hibridación y/o las condiciones de lavado pueden ajustarse para hibridarse con secuencias de la identidad deseada. Por ejemplo, si se buscan secuencias con identidad > 90%, la Tm puede reducirse 10°C. En general, las condiciones de fidelidad se seleccionan de forma que sean aproximadamente 5°C inferiores que el punto de fusión térmico (Tm) para la secuencia específica y su complemento para una resistencia iónica y un pH definidos. Sin embargo, las condiciones de fidelidad muy estrictas pueden usar una hibridación y/o un lavado a 1, 2, 3 o 4°C por debajo del punto de fusión térmico (Tm); las condiciones de fidelidad moderadas pueden usar una hibridación y/o un lavado a 6, 7, 8, 9 ó 10°C por debajo del punto de fusión térmico (Tm); las condiciones de fidelidad bajas pueden usar una hibridación y/o un lavado a 11, 12, 13, 14, 15 ó 20°C por debajo del punto de fusión térmico (Tm). Usando la ecuación, la hibridación y las composiciones de lavado, y la Tm deseada, los expertos en la materia comprenderán que las variaciones en la fidelidad de hibridación y/o las soluciones de lavado están descritas de forma intrínseca. Si el grado deseado de malapareamiento produce una Tm por debajo de 45°C (solución acuosa) o 32°C (solución de formamida) se prefiere aumentar la concentración de SSC de manera que pueda usarse una temperatura más elevada. Puede encontrarse una extensa guía sobre la hibridación de ácido nucleicos en Tijssen, Laboratory Techniques in Biochemistry and Molecular Biology-Hibridysation with Nucleic Acid Probes, Part 1, Chapter 2 "Overview of principles of hybridisation and the strategy of nucleic acid probe assays", Elsevier, N.Y. (1993); y Current Protocols in Molecular Biology, Chapter 2, Ausubel, y col., Eds., Greene Publishing and Wiley-Interscience, Nueva York (1995).

Descripción detallada de la invención

[0041] En un primer aspecto la presente invención se refiere a un procedimiento para genotipificación de marcadores genéticos en una muestra, que comprende las etapas consistentes en:

(a) suministro de una muestra de ADN;

(b) reducción de la complejidad de la muestra de ácido nucleico usando AFLP para producir un subconjunto reproducible de fragmentos de restricción ligados a un adaptador amplificados;

(c) secuenciación del subconjunto reproducible de fragmentos de restricción ligados a un adaptador amplificados usando secuenciación de alto rendimiento;

(d) alineación de las secuencias del subconjunto reproducible de fragmentos de restricción ligados a un adaptador amplificados;

(e) selección de fragmentos de restricción ligados a un adaptador amplificados a partir del subconjunto que ha sido secuenciado al menos 6 veces;

(f) identificación de marcadores genéticos en la selección de fragmentos de restricción ligados a un adaptador amplificados; y

(g) determinación de genotipos codominantes de los marcadores genéticos en la selección de fragmentos de restricción ligados a un adaptador amplificados.

[0042] Por otra parte, la presente descripción desvela también un procedimiento para el descubrimiento, la detección y la genotipificación a gran escala de alto rendimiento de uno o más marcadores genéticos en una o más muestras, que comprende las etapas consistentes en:

(a) suministro de ADN a partir de una o más muestras;

(b) restricción del ADN con al menos una endonucleasa de restricción para producir fragmentos de restricción;

(c) ligación de los adaptadores a los fragmentos de restricción para producir fragmentos de restricción ligados a adaptador;

(d) opcionalmente, amplificación de los fragmentos ligados al adaptador de restricción con un par de cebadores que sea al menos complementario a los adaptadores para producir fragmentos de restricción ligados a un adaptador preamplificados;

(e) amplificación de los fragmentos de restricción ligados a adaptador (opcionalmente preamplificados) con un par de

cebadores, en el que al menos uno de los cebadores contiene una etiqueta identificadora en el extremo 5' del cebador para producir una biblioteca de subconjuntos amplificados etiquetados de fragmentos de restricción ligados a adaptador para cada muestra;

(f) opcionalmente, agrupación de las bibliotecas obtenidas de múltiples muestras;

(g) secuenciación de las bibliotecas usando tecnología de secuenciación de alto rendimiento;

(h) agrupación de las secuencias por biblioteca, usando la etiqueta de identificador;

(i) identificación de los marcadores genéticos comparando secuencias agrupadas en una biblioteca y/o entre las bibliotecas;

(j) determinación de los genotipos (co)dominantes de los marcadores genéticos en la una o más bibliotecas, preferentemente para todas las muestras y para todos los marcadores identificados.

[0043] El procedimiento se refiere al descubrimiento, la detección y la genotipificación de uno o más marcadores genéticos en una o más muestras. En algunas formas de realización, los procedimientos se refieren a la valoración de presencia/ausencia de los marcadores genéticos de interés. En algunas formas de realización el procedimiento se refiere a la determinación de genotipos (co)dominantes de una o más muestras para uno o más marcadores genéticos. Esto puede requerir la normalización del número observado de secuencias de marcadores o alelos de marcadores entre las muestras.

[0044] En la primera etapa (a) del procedimiento debe proporcionarse el ADN. Esto puede realizarse mediante procedimientos conocidos en la técnica de por sí. El aislamiento de ADN se consigue en general usando procedimientos comunes en la técnica como la recogida de tejido de un miembro de la población, la extracción de ADN (por ejemplo usando el kit de ADN Q-Biogene fast), la cuantificación y normalización para obtener cantidades iguales de ADN por muestra. El ADN puede obtenerse de distintas fuentes (Genomic, ARN, ADNc, BAc, YAC etc.) y organismos (humano, mamífero, planta, microorganismos, etc.). El ADN aislado puede guardarse en reserva.

[0045] El ADN se restringe en la etapa (b) usando al menos una endonucleasa de restricción. Según el caso, es decir, el tamaño de genoma, pueden usarse más endonucleasas. En algunas formas de realización, pueden usarse 2 o más endonucleasas. En la mayoría de los genomas son suficientes 2 endonucleasas y por ello este valor es el más preferido. En algunas formas de realización, especialmente para genomas largos o complejos, pueden usarse más endonucleasas. Preferentemente la endonucleasa proporciona fragmentos de restricción relativamente cortos en el orden de 250-500 pb, si bien esto no es esencial. Normalmente, se prefiere al menos una endonucleasa de corte frecuente, es decir, endonucleasas que tienen una secuencia de reconocimiento de 4 ó 5 pares de bases. Una de estas enzimas es MseI, si bien existen otras muchas disponibles comercialmente y pueden usarse. También pueden usarse enzimas que cortan fuera de su secuencia de reconocimiento (tipo IIs), o enzimas que proporcionan fragmentos de restricción de extremos romos. Una combinación preferida usa una secuencia rara (secuencia de reconocimiento de 6 y más pares de bases, por ejemplo EcoRI) y un elemento de corte frecuente.

[0046] Después de la restricción de los ADN en reserva, o simultáneamente, los adaptadores se ligan a los fragmentos de restricción para proporcionar fragmentos de restricción ligados a adaptador. Puede usarse uno o más adaptadores diferentes, por ejemplo dos adaptadores, uno directo y uno inverso. Alternativamente puede usarse un adaptador para todos los fragmentos o conjuntos de adaptadores que en el extremo saliente del adaptador contengan permutaciones de nucleótidos tales que proporcionen ligadores de indexación que pueden permitir una etapa de preselección (Unrau y col., Gene, 1994, 145, 163-169). Alternativamente, pueden usarse adaptadores de extremo romo, en el caso de fragmentos de restricción de extremo romo. La ligación por adaptador es bien conocida en la técnica y se describe entre otros en el documento EP-534.858. Una variante útil de la tecnología de AFLP no usa nucleótidos selectivos (en su caso, cebadores +0/+0) y a veces recibe el nombre de PCR de ligación. Al igual que en PCR, la etapa de selección se proporciona mediante el uso de enzimas de restricción, de manera que diferentes enzimas de restricción producen diferentes subconjuntos. Esto se denota también a veces como preamplificación en la que se usan cebadores que son al menos complementarios a los adaptadores y opcionalmente también a parte de los restos de la secuencia de reconocimiento de la endonucleasa de restricción. La preamplificación puede servir para normalizar (adicionalmente) la cantidad de ADN de cada muestra, o para aumentar la cantidad total de ADN de modo que se permitan múltiples análisis (es decir, división de las muestras) y para mejorar la relación señal-ruido. La preamplificación puede usarse también para introducir etiquetas que permitan el agrupamiento antes de la amplificación selectiva. Mediante la introducción de etiquetas de nucleótidos

(por ejemplo 4 pb) en el extremo 5' del cebador, pueden marcarse los fragmentos de restricción para una muestra distinta y en el extremo del proceso pueden recuperarse usando la etiqueta.

[0047] Los fragmentos de restricción ligados a adaptador son amplificados después de la preamplificación opcional, en la etapa (d) del procedimiento de la invención con un par de cebadores. Uno de los cebadores es complementario a al menos parte del adaptador y puede ser complementario además a parte del resto de la secuencia de reconocimiento de la endonucleasa y puede contener además nucleótidos (seleccionados aleatoriamente) selectivos en su extremo 3', de forma similar a como se describe en el documento EP-534.858. Preferentemente los cebadores son capaces de hibridarse selectivamente en condiciones de fidelidad de hibridación. La amplificación selectiva puede realizarse también con cebadores que llevan una etiqueta en 5' para identificar el origen de la muestra, de forma similar a lo expuesto anteriormente. El resultado es una biblioteca de subconjuntos (etiquetados) de fragmentos de restricción ligados a un adaptador amplificados.

[0048] Los fragmentos amplificados selectivamente en las bibliotecas preparadas a partir de múltiples muestras pueden agruparse opcionalmente en este punto. Esta acción puede ser útil en caso de que se busquen marcadores específicos para ciertos grupos de muestras, como los que comparten determinadas características fenotípicas. El cribado de las muestras agrupadas en reserva suele referirse como análisis de segregación en masa (BSA; Michelmore, Paran y Kesseli, 1991). En algunas formas de realización, la agrupación en reserva también puede realizarse antes de la extracción de ADN en la fase de muestreo, reduciendo el número de preparaciones de ADN. La agrupación en reserva del ADN sirve además para normalizar los ADN antes de la amplificación por PCR para proporcionar una representación más igual en las bibliotecas para secuenciación.

[0049] Las bibliotecas opcionalmente agrupadas de fragmentos de restricción ligados selectivamente a un adaptador amplificados se secuencian actualmente usando tecnología de secuenciación de alto rendimiento.

[0050] En principio, la secuenciación puede realizarse por cualquier medio conocido en la técnica, tal como el procedimiento de terminación de cadenas didesoxi (secuenciación de Sanger). Sin embargo se prefiere y es más ventajoso que la secuenciación se realice usando procedimientos de secuenciación de alto rendimiento, por ejemplo, mediante los procedimientos desvelados en los documentos WO-03/004.690, WO-03/054.142, WO-2004/069.849, WO-2004/070.005, WO-2004/070.007 y WO-2005/003.375 (todos en nombre de 454 Life Sciences), de Seo y col. (2004) Proc. Natl. Acad. Sci. USA 101:5488-93, y tecnologías de Helios, Solexa, US Genomics, etc. Se prefiere sobre todo que la secuenciación se realice usando el aparato y/o el procedimiento desvelados en los documentos WO-03/004.690, WO-03/054.142, WO-2004/069.849, WO-2004/070.005, WO-2004/070.007 y WO-2005/003.375 (todos en nombre de 454 Life Sciences). La tecnología descrita actualmente permite la secuenciación de hasta 40 millones de bases en una única ejecución y es 100 veces más rápida y económica que la tecnología competidora basada en secuenciación de Sanger y que usa instrumentos de electroforesis capilar disponibles actualmente como MegaBACE (GE Healthcare) o ABI3700(xl) (Applied Biosystems). Aumentará al aumentar la longitud de lectura por reacción y/o el número creciente de reacciones en paralelo. La tecnología de secuenciación consiste aproximadamente en 5 etapas: 1) fragmentación de ADN y ligación de adaptador específico para crear una biblioteca de ADN monocatenario (ADNmc); 2) apareamiento de ADNmc con microesferas, emulsificación de las microesferas en microrreactores de agua en aceite y ejecución de PCR de emulsión para amplificar las moléculas de ADNmc individuales en las microesferas; 3) selección/enriquecimiento de microesferas que contienen moléculas de ADNmc amplificado en su superficie, 4) deposición de ADN que contiene microesferas en una PicoTiterPlate®; y 5) secuenciación simultánea en 100.000 pocillos por generación de una señal luminosa de pirofosfato.

[0051] En una realización preferida, la secuenciación comprende las etapas consistentes en:

(1) apareamiento de fragmentos ligados a adaptador de secuenciación a microesferas, cada apareamiento de microesfera con un único fragmento;

(2) emulsificación de las microesferas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una única microesfera;

(3) realización de PCR de emulsión para amplificar los fragmentos ligados a adaptador en la superficie de las microesferas;

(4) selección/enriquecimiento de las microesferas que contienen fragmentos ligados a adaptador amplificados;

(5) carga de las microesferas en pocillos, comprendiendo cada pocillo una única microesfera; y

(6) generación de una señal de pirofosfato.

En la primera etapa (1), los adaptadores que están presentes en los fragmentos de restricción ligados a adaptador se aparean con las microesferas. Tal como se indica en la presente memoria descriptiva anteriormente, el adaptador de secuenciación incluye al menos una región "clave" para el apareamiento con una microesfera, una región de cebador de secuenciación y una región de cebador de PCR. En particular, los fragmentos de restricción ligados a un adaptador amplificados contienen actualmente en uno de los extremos la siguiente secuencia 5'-Sitio de unión de cebador de secuencia---Etiqueta---secuencia de cebador de PCR-3', mientras que en el otro extremo existe un segmento que puede ser como sigue: 5'-Secuencia de apareamiento de microesfera---Etiqueta---Secuencia específica de adaptador---secuencia específica de sitio de restricción (opcional)---secuencia (aleatoriamente) selectiva (opcional)- 3'. Puede estar claro que el Sitio de unión de cebador de secuencia y la Secuencia de apareamiento de microesfera son intercambiables. Esta Secuencia de apareamiento de microesfera puede usarse ahora para el apareamiento de los fragmentos a la microesfera, conteniendo la microesfera una secuencia de nucleótidos en ese extremo.

[0052] Así, los fragmentos adaptados se aparean con microesferas, cada apareamiento con microesfera con un fragmento adaptado individual. A la reserva de fragmentos adaptados se añaden microesferas en exceso para asegurar el apareamiento de un único fragmento adaptado por microesfera para la mayoría de las microesferas (distribución de Poisson).

[0053] En una realización preferida, para aumentar aún más la eficacia del cribado, es beneficioso amplificar el producto de PCR direccionalmente en la microesfera para secuenciación. Esto puede realizarse para realizar la PCR con cebadores de PCR con cola de adaptador entre los cuales una cadena del adaptador en el lado de MseI (u otra enzima de restricción) es complementaria al oligonucleótido acoplado con las microesferas de la secuencia.

[0054] En la etapa siguiente, las microesferas se emulsionan en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una única microesfera. Los reactivos de PCR están presentes en los microrreactores de agua en aceite permitiendo que tenga lugar una reacción de PCR en los microrreactores. Posteriormente, los microrreactores se descomponen, y las microesferas que comprenden ADN (microesferas positivas en ADN) quedan enriquecidas.

[0055] En la siguiente etapa, las microesferas se cargan en pocillos, comprendiendo cada pocillo una única microesfera. Los pocillos forman parte preferentemente de una PicoTiterPlate™ que permite la secuenciación simultánea de una gran cantidad de fragmentos.

[0056] Después de la adición de microesferas portadoras de enzimas, la secuencia de los fragmentos se determina usando pirosecuenciación. En etapas sucesivas, la PicoTiterPlate™ y las microesferas así como las microesferas con enzimas se someten a diferentes desoxirribonucleótidos en presencia de reactivos de secuenciación convencionales, y tras la incorporación de un desoxirribonucleótido se genera una señal luminosa que se registra. La incorporación del nucleótido correcto generará una señal de pirosecuenciación que puede detectarse.

[0057] La pirosecuenciación en sí es conocida en la técnica y ha sido descrita entre otros en www.biotagebio.com; en la sección de tecnología de www.pyrosequencing.com/. La tecnología se aplica además, por ejemplo, en los documentos WO-03/004690, WO-03/054.142, WO-2004/069.849, WO-2004/070.005, WO-2004/070.007 y WO-2005/003.375 (todos en nombre de 454 Life Sciences).

[0058] Después de la secuenciación, las secuencias de los fragmentos que se obtienen directamente de la etapa de secuenciación pueden recortarse, preferentemente *in silico,* para eliminar cualquier secuencia de apareamiento de microesferas, cebador de secuenciación, adaptador o información relacionada con secuencias de cebador.

[0059] Normalmente, la alineación o agrupación se realiza en datos de secuencias que han sido recortador para cualquier secuencia añadida de adaptadores/cebador, es decir, usando sólo los datos de secuencias de los fragmentos que proceden de la muestra de ácido nucleico, junto con la etiqueta de identificador opcional.

[0060] Los procedimientos de alineación de secuencias para comparación son bien conocidos en la técnica. Se han descrito varios programas y algoritmos de alineación en: Smith y Waterman (1981) Adv. Appl. Math. 2:482; Needleman y Wunsch (1970) J. Mol. Biol. 48:443; Pearson y Lipman (1988) Proc. Natl. Acad. Sci. USA 85:2444;

Higgins y Sharp (1988) Gene 73:237-244; Higgins y Sharp (1989) CABIOS 5:151-153; Corpet y col. (1988) Nucl. Acids Res. 16:10881-90; Huang y col. (1992) Computer Appl. in the Biosci. 8:155-65; y Pearson y col. (1994) Meth. Mol. Biol. 24:307-31. Altschul y col. (1994) Nature Genet. 6:119-29 presentan una consideración detallada de procedimientos de alineación de secuencias y cálculos de homología.

[0061] La herramienta de búsqueda de alineación local básica NCBI (BLAST) (Altschul y col., 1990) está disponible en varias fuentes, entre ellas el National Center for Biological Information (NCBI, Bethesda, Md.) y en Internet, para su uso en relación con los programas de análisis de secuencias blastp, blastn, blastx, tblastn y tblastx. Puede accederse a ella en <http://www.ncbi. nlm.nih.gov/BLAST/>. Una descripción de cómo determinar la identidad de secuencias usando este programa está disponible en <http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html>. La base de datos comprende preferentemente secuencias EST, secuencias genómicas de las especies de interés y/o la base de secuencias no redundante de GenBank o bases de datos de secuencias similares.

[0062] Los procedimientos de secuenciación de alto rendimiento pueden usarse tal como se describe en Shendure y col. Science, Vol 309, Issue 5741, 1728-1732. Algunos ejemplos de los mismos son secuenciación microelectroforética, secuenciación de hibridación/secuenciación por hibridación (SBH), secuenciación de matriz cíclica en moléculas amplificadas, secuenciación de matriz cíclica en moléculas individuales, procedimientos en tiempo real no cíclicos en moléculas individuales, tales como secuenciación de polimerasas, secuenciación de exonucleasas, secuenciación de nanoporos.

[0063] En la biblioteca puede determinarse ahora la presencia de marcadores genéticos y/o el genotipo de la muestra para marcadores genéticos.

[0064] El procedimiento de la presente invención puede usarse para la identificación, detección de marcadores de AFLP y determinación de genotipos, pero también para la identificación, detección y genotipificación de marcadores de SNP contenidos en bandas constantes. Para proporcionar una solución al problema de variación del muestreo que afecta a la precisión de la genotipificación de marcadores genéticos por fragmentos alélicos (de marcadores) de secuenciación contenidos en una biblioteca de múltiples fragmentos, los autores de la presente invención han encontrado también que la detección de marcadores de AFLP por medio de secuenciación se realiza preferentemente con suficiente redundancia (profundidad) para muestrear todos los fragmentos amplificados al menos una vez y acompañado de medios estadísticos que abordan la cuestión de la variación del muestreo en relación con la exactitud de los genotipos invocados. Además, al igual que con la valoración de AFLP, en el contexto de una población segregada, la valoración simultánea de los progenitores en un experimento ayudará a determinar el umbral estadístico, dado que todos los posibles alelos de la muestra se valorarán como progenitor 1 o progenitor 2. Debe observarse que se sugiere muestrear los progenitores con mayor redundancia que los individuos de las poblaciones segregadas.

[0065] Así, en algunas formas de realización, la redundancia de los fragmentos de restricción etiquetados ligados a un adaptador amplificados es de al menos 6, preferentemente al menos 7, más preferentemente al menos 8 y con la máxima preferencia al menos 9. En algunas formas de realización, la secuencia de cada fragmento de restricción ligada a adaptador se determina al menos 6, preferentemente al menos 7, más preferentemente al menos 8 y con la máxima preferencia al menos 9 veces. En algunas formas de realización, la redundancia se selecciona de tal forma que, suponiendo una probabilidad global de 50/50 de identificar el locus correctamente como homocigoto, la probabilidad de una identificación correcta del locus sea de más del 95%, 96%, 97%, 98%, 99%, 99,5%.

[0066] En algunas formas de realización, el número de muestras puede estar comprendido entre 1 y 100.000, lo que depende también en gran medida del tamaño del genoma que se analizará y del número de fragmentos amplificados selectivamente. Normalmente, la capacidad de la tecnología de secuenciación empleada proporciona el factor más limitante a este respecto.

Breve descripción de los dibujos

[0067]

La **Figura 1A** muestra un fragmento según la presente invención apareado en una microesfera ('microesfera 454') y la secuencia de cebador usada para preamplificación de las dos líneas de pimiento. 'Fragmento de ADN' denota el fragmento obtenido después de la digestión con una endonucleasa de restricción, 'adaptador keygene' denota un adaptador que proporciona un sitio de apareamiento para los cebadores de oligonucleótidos (fosforilados) usados para generar una biblioteca, 'KRS' denota una secuencia de identificador (etiqueta), 'adaptador SEQ 454' denota un

adaptador de secuenciación, y 'adaptador de PCR 454' denota un adaptador para permitir la amplificación por emulsión del fragmento de ADN. El adaptador de PCR permite el apareamiento con la microesfera y la amplificación y puede contener un saliente en 3'.

La **Figura 1B** muestra un cebador esquemático usado en la etapa de reducción de la complejidad. Dicho cebador comprende en general una región de sitio de reconocimiento indicada como (2), una región constante que puede incluir una sección de etiqueta indicada como (1) y uno o más nucleótidos selectivos en una región selectiva indicada como (3) en el extremo 3' de los mismos.

La **Figura 2** muestra la estimación de concentración de ADN usando electroforesis en gel de agarosa al 2%. S1 denota PSP11; S2 denota PI201234. 50, 100, 250 y 500 ng denotan respectivamente 50 ng, 100 ng, 250 ng y 500 ng para estimar las cantidades de ADN de S1 y S2. Las Fig. 2C y 2D muestran la determinación de la concentración de ADN usando espectrofotometría Nanodrop.

La **Figura 3** muestra los resultados de las valoraciones intermedias de calidad del ejemplo 3.

La **Figura 4** muestra los organigramas del desarrollo del procesamiento de datos de secuencias, es decir, las etapas adoptadas desde la generación de los datos de secuenciación a la identificación de posibles SNP, SSR e indeles, a través de las etapas de la eliminación de la información de secuencias conocida en Recorte y Etiquetado que produce datos de secuencias recortadas que se agrupan y ensamblan para producir cóntigos y singletones (fragmentos que no pueden ensamblarse en un cóntigo) después de que puedan identificarse y valorarse los posibles polimorfismos. La Figura 4B elabora adicionalmente el proceso de investigación de polimorfismos.

La **Figura 5** es una alineación múltiple "10037_CL989contig2" de secuencias de fragmentos de AFLP de pimiento, que contienen un posible polimorfismo de nucleótido simple (SNP). Obsérvese que el SNP (indicado por una flecha negra) se define por un alelo A presente en las dos lecturas de la muestra 1 (PSP11), denotado por la presencia de la etiqueta MS1 en el nombre de las dos lecturas superiores, y un alelo G presente en la muestra 2 (PI201234), denotado por la presencia de la etiqueta MS2 en el nombre de las dos lecturas inferiores. Los nombres de las lecturas se muestran a la izquierda. La secuencia de consenso de esta alineación múltiple es (5'- 3'):

TAACACGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACA[A/G]TGTTGGTTTT

GGTGCTAACTTCAACCCCACTACTGTTTTGCTCTATTTTTG.

La **Figura 6** es una representación gráfica de la probabilidad de clasificación correcta del genotipo basándose en el número de lecturas observadas por locus.

Ejemplos

[0068] El procedimiento se ilustra del modo siguiente:

1) Se preparan plantillas AFLP según un protocolo modificado de Vos *y col.* que incluye una etapa de desnaturalización por calor durante 20 min a 80°C entre las etapas de restricción y ligación. Después de incubación durante 20 min a 80°C, se enfría el digesto de la enzima de restricción a temperatura ambiente y se añade ADN ligasa. La etapa de desnaturalización conduce a disociación de las cadenas complementarias de los fragmentos de restricción hasta 120 pb de manera que no se ligarán adaptadores a los extremos. En consecuencia, los fragmentos inferiores a 120 pb no se amplificarán, con lo que se consigue una selección de tamaño.

2) Las reacciones de preamplificación, si fueran aplicables, se realizan como en AFLP convencional.

3) La última etapa de amplificación (selectiva) se realiza usando cebadores AFLP con etiquetas de identificador únicas para cada muestra en la población/experimento (usando una única secuencia de identificador de 4 pb; KIS). Los KIS están situados en el extremo 5' de los cebadores AFLP selectivos. Se usará un nucleótido selectivo adicional en comparación con el número de bases selectivas usadas en detección de AFLP convencional por electroforesis, por ejemplo +4/+3 para una huella genética EcoRI/MseI en pimiento (detección de gel +3/+3) y +4/+4 para huella genética EcoRI/MseI en maíz (detección de gel +4/+3). El número de nucleótidos selectivos que se aplican debe determinarse empíricamente; puede ser tal que puede aplicarse el mismo número de nucleótidos selectivos que se usa en la detección de gel. Este número depende además del número de muestras incluidas en el experimento, dado que se supone que el número de trazas de secuencias se fija en 200.000 en el estado actual de

la tecnología de secuenciación, aunque puede aumentar y probablemente lo hará. El punto de partida preferido consiste en conseguir un muestreo de 10 veces de fragmentos de AFLP por biblioteca de muestras.

4) La recogida de muestras preparadas según las etapas 1-4 se somete a secuenciación por medio de la tecnología 454 Life Sciences. Esto significa que se clonan fragmentos de AFLP individuales en microesferas, se amplifican por PCR y se secuencian. Se espera una producción de 200.000 secuencias de 100 pb de longitud. Para una recogida de 100 muestras, esto equivale a un promedio de 2.000 trazas de secuencias/muestra, trazable hasta el nº de muestra por medio de la etiqueta en 5'.

5) Suponiendo la amplificación de 100 fragmentos de AFLP por PC cuando se usa 1 nucleótido selectivo adicional en comparación con el número usado con detección de gel, en el que el 90% son bandas constantes, los fragmentos de AFLP son muestreados con una redundancia media de 20 veces por fragmento. Sin embargo, dado que la secuenciación es no direccional y la mayoría de las bandas son > 200 pb, la redundancia de secuenciación será ligeramente superior a 10 veces para cada extremo de fragmento.

6) Todas las secuencias se agrupan por muestra usando la etiqueta KRS. Dado un sobremuestreo de 10 veces, esto significa que se esperan 200 trazas de secuencias diferentes por muestra, lo que representa 200 x 100 pb = 20 kb secuencia/muestra. Cuando el 10% de estas secuencias proceden de marcadores de AFLP (es decir, 1 alelo se amplifica y el otro está ausente en la reacción PCR), el 90% (18 kb) de las secuencias proceden de bandas constantes.

7) Se valoran dos tipos de marcadores genéticos:

A) Marcadores de AFLP: son secuencias que se observan en algunas muestras, pero están ausentes en otras. La inspección de la frecuencia de secuencias en la recogida de muestras revelará esta categoría. La valoración dominante se realiza dependiendo de la presencia/ausencia de observación de estas secuencias en cada muestra. Una valoración fiable de marcadores de AFLP requiere la fijación de un umbral estadístico en relación con la frecuencia con la que se observan otras secuencias de AFLP en el experimento. Es decir, un marcador de AFLP puede valorarse como presente (dominante) si el marcador de secuencia de AFLP se observa en la muestra, pero la fiabilidad de la valoración ausente depende de la frecuencia (media) de fragmentos de AFLP (constantes). Se requieren niveles de umbral estadístico de manera que se realice la valoración de presencia/ausencia con preferentemente al menos el 99,5% de exactitud, dependiendo del nivel aceptable necesario para la aplicación específica. Si se analiza una población segregada y sus progenitores, estos marcadores pueden valorarse posiblemente de forma codominante así como definir categorías de frecuencias de las secuencias de marcador. Esto último puede ser complicado en la práctica por la influencia de variación de muestreo del marcador de AFLP que difiere entre las muestras.

B) Polimorfismos (SN) en fragmentos de AFLP constantes.

Esta es la categoría más interesante (y abundante) de marcadores genéticos. La esencia es que los marcadores de SNP contenidos en las secuencias internas de fragmentos de AFLP constantes se valoran como marcadores de SNP codominantes. De nuevo, esto requiere preferentemente la aplicación de un nivel de umbral estadístico para la invocación precisa de la presencia o ausencia de un alelo. Se espera que una redundancia de secuenciación de 10 veces de la biblioteca de fragmentos sea suficiente aunque se necesita un análisis estadístico para determinar la exactitud de los genotipos de marcadores de SNP dependiendo del número de cada secuencia de alelos que se observe. La argumentación es que cuando una banda constante contiene un SNP y se observa un alelo, por ejemplo, 5 veces mientras (la secuencia que contiene) el otro alelo no se observa, es muy probable que la muestra sea homocigota para el alelo observado. En consecuencia, cuando se observan los dos alelos, la muestra se valora como heterocigota para el marcador de SNP, con independencia de sus frecuencias.

8) El resultado será una genotipificación de la tabla que contiene los genotipos de marcadores de AFLP valorados de forma (co)dominante y SNP valorados de forma codominante, junto con probabilidades para la corrección de los genotipos de todos los marcadores. Alternativamente, se genera un conjunto de datos que contiene genotipos que han sobrepasado el nivel de umbral estadístico establecido.

[0069] Este enfoque supone un sobremuestreo de 10 veces de fragmentos de AFLP por muestra, lo que produce 18 kb de secuencia/muestra constante y 2 kb de secuencias de marcadores de AFLP.

[0070] El número de marcadores genéticos observados depende de la tasa de SNP en el plasma germinal

investigado. A continuación se suministran estimaciones de los números de marcadores genéticos para diferentes tasas de SNP en plasma germinal, cuando se muestrean secuencias de 20 kb. Se supone que la longitud media de los marcadores/fragmentos de AFLP es 200 pb:

Tabla 1. Números esperados de marcadores genéticos valorados por secuenciación de fragmentos de AFLP usando

Tecnología 454 Life sciences suponiendo sobremuestreo de 10 veces, 200.000 trazas de secuencias, 90% de bandas constantes/10% de marcadores de AFLP para diversas tasas de SNP.		
Tasa de SNP	Marcadores de AFLP (2 kb)	SNP en bandas constantes (18 kb) *
1/250 pb	8	72
1/1.000 pb	2	18
1/2.000 pb	1	9
1/5.000 pb	0,4	3,6
* Dado que los fragmentos de AFLP pueden secuenciarse desde los dos extremos, una proporción de los SNP observados puede proceder de los mismos loci.		

[0071] Es importante observar que los números suministrados en la tabla 1 son promedios, que pueden diferir entre combinaciones de distintos cebadores. De forma análoga a la tipificación de AFLP convencional, la identificación de las combinaciones de cebadores superiores (PC) puede producir números más elevados de marcadores por PC. Además, los números presentados en la Tabla 1 pueden cambiar dependiendo del nivel requerido de sobremuestreo necesario para alcanzar el nivel de exactitud exigido.

[0072] El cálculo de la clasificación correcta del genotipo es el siguiente:

$$P\ (correcto) = P\ (aa) + P\ (AA) + P\ (Aa) * [1 - 0{,}5 * \exp(n - 1)]$$

[0073] En el que P(aa) es la fracción de la población con genotipo aa (en el gráfico adjunto, fig. 9, fijado en 0,25). P(AA) es la fracción de la población con genotipo AA (fijado en 0,25). P(Aa) es la fracción de la población con genotipo Aa (en fig. 6 y tabla mostrada a continuación, fijado en 0,5). n es igual el número de sujetos.

Tabla

n	P
1	0,5
2	0,75
3	0,875
4	0,9375
5	0,96875
6	0,984375
7	0,992188
8	0,996094
9	0,998047
10	0,999023

Ejemplo 1 PIMIENTO

[0074] Se usó el ADN de las líneas de *Pimiento* PSP-11 y PI201234 para generar un producto de AFLP mediante el uso de cebadores específicos del *Sitio de reconocimiento Keygene* AFLP. (Estos cebadores de AFLP son esencialmente los mismos que los cebadores de AFLP convencionales, descritos por ejemplo en el documento EP-0.534.858, y en general contendrán una *región de sitio de reconocimiento,* una región constante y uno o más nucleótidos selectivos en una región selectiva).

[0075] A partir de las líneas de pimiento PSP-11 o PI201234 se digirieron 150 ng de ADN con las endonucleasas de restricción *EcoRI* (5U/reacción) y *MseI* (2U/reacción) durante 1 hora a 37°C seguido por inactivación durante 10 minutos a 80°C. Los fragmentos de restricción obtenidos se ligaron con adaptador de oligonucleótidos sintéticos bicatenarios, un extremo del cual es compatible con uno o con los dos extremos de los fragmentos de restricción EcoRI y/o MseI. La mezcla de ligación de restricción se diluyó 10 veces y se preamplificaron 5 microlitros de cada muestra (2) con cebadores EcoRI +1(A) y MseI +1(C) (conjunto I). Después de amplificación se verificó la calidad del producto de preamplificación de las dos muestras de pimiento en un gel de agarosa al 1%. Los productos de preamplificación se diluyeron 20 veces, seguido por preamplificación AFLP

KRSEcoRI +1(A) y KRSMseI +2(CA). Las secciones (de identificador) de KRS están subrayadas y los nucleótidos selectivos se marcan en negrita en el extremo 3' en la secuencia de cebador SEQ ID 1-4 mostrada a continuación. Después de la amplificación se verificó la calidad del producto de preamplificación de las dos muestras de pimiento en un gel de agarosa al 1% y con una huella genética de AFLP EcoRI +3(A) y MseI +3(C) (3) (4). Se purificaron por separado los productos de preamplificación de las dos líneas de pimiento en una columna de QiagenPCR (5). Se midió la concentración de las muestras en un espectrofotómetro NanoDrop® ND-1000. Se mezcló un total de 5 microgramos de PSP-11 y 5 microgramos de PI201234 de productos de PCR y se secuenciaron.

[0076] Conjunto de cebador I usado para preamplificación de PSP-11

E01LKRS1 5'-CGTCAGACTGCGTACCAATTC**A**-3' [SEQ ID 1]

M15KKRS1 5'-TGGTGATGAGTCCTGAGTAA**CA**-3' [SEQ ID 2]

[0077] Conjunto de cebador II usado para preamplificación de PI201234

E01LKRS2 5'-CAAGAGACTGCGTACCAATTC**A**-3' [SEQ ID 3]

M15KKRS2 5'-AGCCGATGAGTCCTGAGTAA**CA**-3' [SEQ ID 4]

(1) EcoRI/MseI restricción ligación mezcla

Mezcla de restricción (40 µl/muestra)

[0078]

ADN	6 µl (6300 ng)
ECoRI (5U)	0,1 µl
MseI (2U)	0,05 µl
5xRL	8 µl
MQ	25,85 µl
Total	40 µl

Incubación durante 1 h a 37°C

Adición de:

Mezcla de ligación (10 µl/muestra)

[0079]

ATP 10 mM	1 µl
T4 ADN ligasa	1 µl
Adapt. ECoRI. (5 pmol/µl)	1 µl
Adapt. MseI. (50 pmol/µl)	1 µl
5xRL	2 µl
MQ	4 µl
Total	10 µl

Incubación durante 3 h a 37°C

Adaptador EcoRI

[0080]

91M35/91M36:	*-CTCGTAGACTGCGTACC	:91M35 [SEQ ID 5]
± bio	CATCTGACGCATGGTTAA	:91M36 [SEQ ID 6]

Adaptador MseI

[0081]

92A18/92A19: 5-GACGATGAGTCCTGAG-3 :92A18 [SEQ ID 7]
3-TACTCAGGACTCAT-5 :92A19 [SEQ ID 8]

(2) Preamplificación

Preamplificación (A/C):

[0082]

Mezcla RL (10x)	5 µl
EcoRI-pr E01L (50 ng/µl)	0,6 µl
MseI-pr M02K (50 ng/µl)	0,6 µl
dNTPs (25 mM)	0,16 µl
Taq.pol. (5U)	0,08 µl
10XPCR	2,0 µl
MQ	11,56 µl
Total	20 µl/reacción

Perfil térmico de preamplificación

[0083] Se realizó la preamplificación selectiva en un volumen de reacción de 50 µl. La PCR se llevó a cabo en un PE GeneAmp PCR System 9700 y se inició un perfil de **20** ciclos con una etapa de desnaturalización a 94°C durante 30 segundos, seguido por una etapa de apareamiento de 56°C durante 60 segundos y una etapa de extensión de 72°C durante 60 segundos.

EcoRI +1(A)[1]

[0084]

E01 L 92R11: 5-AGACTGCGTACCAATTCA-3 [SEQ ID 9]

MseI +1 (C)[1]

[0085]

M02k 93E42: 5-GATGAGTCCTGAGTAAC-3 [SEQ ID 10]

Preamplificación A/CA:

[0086]

PA+1/+1-mix (20x)	:5 µl
EcoRI-pr	:1,5 µl
MseI-pr.	:1,5 µl
dNTPs (25 mM)	:0,4 µl
Taq.pol.(5U)	:0,2 µl
10XPCR	:5 µl
MQ	:36,3 µl
Total	: 50 µl

[0087] Se realizó la preamplificación selectiva en un volumen de reacción de 50 µl. La PCR se llevó a cabo en un PE GeneAmp PCR System 9700 y se inició un perfil de 30 ciclos con una etapa de desnaturalización a 94°C durante 30 segundos, seguido por una etapa de apareamiento de 56°C durante 60 segundos y una etapa de extensión de 72°C durante 60 segundos.

(3) KRSEcoRI +1 (A) y KRSMseI +2(CA)^2

[0088]

05F212	E01LKRS1	CGTCAGACTGCGTACCAATTC**A**	-3' [SEQ ID 11]
05F213	E01LKRS2	CAAGAGACTGCGTACCAATTC**A**	-3' [SEQ ID 12]
05F214	M15KKRS1	TGGTGATGAGTCCTGAGTAA**CA**	-3' [SEQ ID 13]
05F215	M15KKRS2	AGCCGATGAGTCCTGAGTAA**CA**	-3' [SEQ ID 14]

nucleótidos selectivos en negrita y etiquetas (KRS) subrayadas

Muestra PSP11 : E01LKR**S1**/M15KKR**S1**
Muestra PI120234 : E01LKR**S2**/M15KKR**S2**

(4) Protocolo AFLP

[0089] Se realizó una amplificación selectiva en un volumen de reacción de 20 µl. La PCR se llevó a cabo en un PE GeneAmp PCR System 9700. Se inició un perfil de **13** ciclos con una etapa de desnaturalización a 94°C durante 30 segundos, seguido por una etapa de apareamiento de 65°C durante 30 segundos, con una fase de descenso en la que la temperatura de apareamiento se redujo 0,7°C en cada ciclo, y una etapa de extensión de 72°C durante 60 segundos. Este perfil se siguió de un perfil de 23 ciclos con una etapa de desnaturalización a 94°C durante 30 segundos, seguido por una etapa de apareamiento de 56°C durante 30 segundos y una etapa de extensión de 72°C durante 60 segundos.

EcoRI +3(**AAC**) y MseI +3(**CAG**)

[0090]

(5) Columna Qiagen

E32	92S02:	5-GACTGCGTACCAATTC**AAC**-3 [SEQ ID 15]
M49	92G23:	5-GATGAGTCCTGAGTAA**CAG**-3 [SEQ ID 16]

[0091] Se purificó el producto de AFLP usando el kit de purificación QIAquick PCR (QIAGEN) de acuerdo con QIAquick® Spin Handbook 07/2002 página 18 y se midió la concentración con un espectrofotómetro NanoDrop® ND-1000. Se reunió un total de 5 µg de producto de AFLP +1/+2 PSP-11 y 5 µg de +1/+2 producto de AFLP PI201234 y se disolvió en 23,3 µl de TE. Finalmente se obtuvo una mezcla con una concentración de 430 ng/µl producto de AFLP +1/+2.

Preparación de biblioteca de secuencias y secuenciación de alto rendimiento

[0092] Los productos de amplificación combinados de las los líneas de pimiento se sometieron a secuenciación de alto rendimiento usando la tecnología de secuenciación 454 Life Sciences tal como se describe en Margulies y col., (Margulies y col., Nature 437, pp. 376-380 y Online Supplements). Específicamente, los productos de PCR AFLP se pulieron en un primer extremo y posteriormente se ligaron con adaptadores para facilitar la amplificación por PCR-emulsión y la posterior secuenciación de fragmentos tal como se describe en Margulies y col. Las secuencias de adaptador 454, los cebadores de PCR-emulsión, los cebadores de secuencias y las condiciones de realización de las secuencias se describieron en Margulies y col. El orden lineal de elementos funcionales en un fragmento de PCR-emulsión amplificado en microesferas de Sefarosa en el proceso de secuenciación 454 fue el siguiente, tal como se ilustra en la figura 1A:

adaptador de PCR 454 - 454 adaptador de secuencias – etiqueta de cebador AFLP de 4 pb 1 – secuencia de cebador de AFLP 1 que incluye nucleótido(s) selectivo(s) – secuencia interna de fragmento de AFLP – secuencia de cebador de AFLP 2 que incluye nucleótido(s) selectivo(s), etiqueta de cebadores AFLP de 4 pb 2 - adaptador de secuencias 454 - adaptador de PCR 454 – microesfera de Sefarosa

[0093] Se realizaron dos ejecuciones de secuencias 454 de alto rendimiento mediante 454 Life Sciences (Branford, CT; Estados Unidos).

Procesamiento de datos de ejecución de secuencias 454.

[0094] Se procesaron los datos de secuencias resultantes de una ejecución de secuencia 454 usando un desarrollo de bioinformática (Keygene N.V.). Específicamente, se convirtieron lecturas de secuencias de invocación de bases 454 en formato FASTA y se inspeccionó la presencia de secuencias de adaptador de AFLP etiquetadas usando un algoritmo BLAST. Tras las correspondencias de alto grado de confianza con las secuencias de cebadores de AFLP etiquetadas, se recortaron las secuencias, se restauraron las endonucleasas de sitios de restricción y se asignaron las etiquetas apropiadas (muestra 1 EcoRI (ES1), muestra 1 MseI (MS1), muestra 2 EcoRI (ES2) o muestra 2 MseI (MS2), respectivamente). A continuación, se agruparon todas las secuencias recortadas de más de 33 bases usando un procedimiento megaBLAST basándose en homologías de secuencias globales. A continuación, se ensamblaron los grupos en uno o más cóntigos y/o singletones por grupo, usando un algoritmo de alineación múltiple CAP3. Se inspeccionaron los cóntigos que contenían más de una secuencia en busca de malapareamientos en las secuencias, que representaban posibles polimorfismos. Se asignó a los malapareamientos de las secuencias una valoración de calidad basada en los siguientes criterios:

* el número de lecturas en un cóntigo

* la distribución de alelos observada

[0095] Los dos criterios anteriores forman la base de la denominada valoración Q asignada a cada posible SNP/indel. Las valoraciones Q están comprendidas entre 0 y 1; una valoración Q de 0,3 sólo puede alcanzarse en el caso de que los dos alelos se observen al menos dos veces.

* posición en homopolímeros de una cierta longitud (ajustable; ajuste por omisión para evitar el polimorfismo localizado en homopolímeros de 3 bases o más).

* número de cóntigos en el grupo.

* distancia hasta los malapareamientos contiguos más próximas (ajustable; importante para ciertos tipos de genotipificación de ensayos que investigan las secuencias de flanqueo)

* el nivel de asociación de alelos observados con la muestra 1 o la muestra 2; en caso de una asociación perfecta y coherente entre los alelos de un posible polimorfismo y las muestras 1 y 2, el polimorfismo (SNP) se indica como un posible polimorfismo de "elite" (SNP). Se piensa que un polimorfismo de elite tiene una alta probabilidad de estar situado en una secuencia genómica única o de baja copia en caso de que se hayan usado dos líneas homocigóticas en el proceso de descubrimiento. Por el contrario, una asociación débil de un polimorfismo con el origen de la muestra comporta un alto riesgo de haber descubierto un falso polimorfismos procedente de la alineación de secuencias no alélicas en un cóntigo.

[0096] Las secuencias que contenían motivos de SSR se identificaron usando la herramienta de búsqueda MISA (herramienta de identificación MIcroSAtellelite; disponible en http://pgrc.ipk-gatersleben.de/misa/).

[0097] En la Tabla mostrada a continuación se recogen las estadísticas globales del proceso.

Tabla. Estadísticas globales de un proceso de secuencia 454 para descubrimiento de SNP en pimiento.

Combinación de enzimas	Ejecución
Recorte	
Todas las lecturas	254308
Defectuosas	5293 (2 %)
Correctas	249015 (98%)
Concatámeros	2156 (8,5 %)
Etiquetas mixtas	1120 (0,4 %)
Lecturas correctas	
Un extremo recortado	240817 (97%)
Dos extremos recortados	8198 (3 %)
Número de lecturas muestra 1	136990 (55%)
Número de lecturas muestra 2	112025 (45 %)
Agrupación	
Número de cóntigos	21918
Lecturas en cóntigos	190861
Número medio de lecturas por cóntigo	8,7
Investigación de SNP	
SNP con valoración Q ≥ 0,3 *	1483
Indel con valoración Q ≥ 0,3 *	3300
Número total de motivos de SSR identificados	359
Número de lecturas que contienen uno o más motivos de SSR	353
Número de motivos de SSR con tamaño de unidad 1 (homopolímero)	0
Número de motivos de SSR con tamaño de unidad 2	102
Número de motivos de SSR con tamaño de unidad 3	240
Número de motivos de SSR con tamaño de unidad 4	17
* Los criterios de investigación de SNP/indel fueron los siguientes:	

[0098] Sin polimorfismos adyacentes con valoración Q superior a 0,1 en las 12 bases de cada lado, no presentes en los homopolímeros de 3 o más bases. Los criterios de investigación no tuvieron en cuenta la asociación consistente con la muestra 1 y 2, es decir, los SNP y los indeles no son necesariamente posibles SNP/indeles de elite

[0099] En la Figura 5 se muestra un ejemplo de una alineación múltiple que contiene un posible polimorfismo de elite simple.

Ejemplo 2 : Maíz

[0100] Se usó ADN de las líneas de *Maíz* B73 y M017 para generar el producto de AFLP mediante el uso de cebadores específicos del *Sitio de reconocimiento Keygene* de AFLP. (Estos cebadores de AFLP son esencialmente los mismos que los cebadores de AFLP convencionales, descritos por ejemplo en el documento EP-0.534.858, y en general contendrán una *región de sitio de reconocimiento,* una región constante y uno o más nucleótidos selectivos en el extremo 3' de los mismos).

Se digirió el ADN de las líneas de pimiento B73 o M017 con las endonucleasas de restricción *TaqI* (5U/reacción) durante 1 hora a 65°C y *MseI* (2U/reacción) durante 1 hora a 37°C seguido por inactivación durante 10 minutos a 80°C. Se ligaron los fragmentos de restricción obtenidos con adaptador de oligonucleótidos sintéticos bicatenarios, un extremo del cual es compatible con uno o los dos extremos de los fragmentos de restricción TaqI y/o MseI.

Se realizaron las reacciones de preamplificación de AFLP (20 µl/reacción) con cebadores de AFLP +1/+1 en una mezcla de restricción-ligación diluida 10 veces. Perfil de PCR 20*(30 s a 94°C + 60 s a 56°C + 120 s a 72°C). Se realizaron las reacciones de AFLP adicionales (50 µl/reacción) con diferentes cebadores de Sitios de reconocimiento Keygene AFLP +2 *TaqI* y *MseI* (tabla mostrada a continuación, las etiquetas están en negritas, los nucleótidos selectivos están subrayados) en producto de preamplificación AFLP +1/+1 TaqI/MseI diluida 20 veces. Perfil PCR: 30* (30 s a 94°C + 60 s a 56°C + 120 s a 72°C). Se purificó el producto de AFLP usando el kit de purificación de PCR QIAquick (QIAGEN) de acuerdo con QIAquick® Spin Handbook 07/2002 página 18 y se midió la concentración

con un espectrofotómetro NanoDrop® ND-1000. Se reunió un total de 1,25 μg de cada producto de AFLP B73 +2/+2 diferente y 1,25 μg de cada producto de AFLP M017 +2/+2 AFLP diferente y se disolvió en 30 μl de TE. Finalmente se obtuvo una mezcla con una concentración de 333 ng/μl de producto de AFLP +2/+2 AFLP.

Tabla

SEQ ID	Cebador de PCR	Secuencia de cebadores	Maíz	Reacción AFLP
[SEQ ID 17]	05G360	**ACGT**GTAGACTGCGTACCGAAA	B73	1
[SEQ ID 18]	05G368	**ACGT**GATGAGTCCTGAGTAACA	B73	1
[SEQ ID 19]	05G362	**CGTA**GTAGACTGCGTACCGAAC	B73	2
[SEQ ID 20]	05G370	**CGTA**GATGAGTCCTGAGTAACA	B73	2
[SEQ ID 21]	05G364	**GTAC**GTAGACTGCGTACCGAAG	B73	3
[SEQ ID 22]	05G372	**GTAC**GATGAGTCCTGAGTAACA	B73	3
[SEQ ID 23]	05G366	**TACG**GTAGACTGCGTACCGAAT	B73	4
[SEQ ID 24]	05G374	**TACG**GATGAGTCCTGAGTAACA	B73	4
[SEQ ID 25]	05G361	**AGTC**GTAGACTGCGTACCGAAA	M017	5
[SEQ ID 26]	05G369	**AGTC**GATGAGTCCTGAGTAACA	M017	5
[SEQ ID 27]	05G363	**CATG**GTAGACTGCGTACCGAAC	M017	6
[SEQ ID 28]	05G371	**CATG**GATGAGTCCTGAGTAACA	M017	6
[SEQ ID 29]	05G365	**GAGC**GTAGACTGCGTACCGAAG	M017	7
[SEQ ID 30]	05G373	**GAGC**GATGAGTCCTGAGTAACA	M017	7
[SEQ ID 31]	05G367	**TGAT**GTAGACTGCGTACCGAAT	M017	8
[SEQ ID 32]	05G375	**TGAT**GATGAGTCCTGAGTAACA	M017	8

[0101] Finalmente se agruparon 4 muestras P1 y 4 muestras P2 y se concentraron. Se obtuvo una cantidad total de 25 μl de producto de ADN y una concentración final de 400 ng/μl (total de 10 μg). En la Figura 3 se muestran las valoraciones de calidad intermedias.

SECUENCIACIÓN POR 454

[0102] Las muestras de fragmentos de AFLP de pimiento y maíz según se ha preparado tal como se ha descrito anteriormente en la presente memoria descriptiva fueron procesadas por 454 Life Sciences tal como se ha descrito (Margulies y col., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437 (7057):376-80. Epub July 31, 2005).

PROCESAMIENTO DE DATOS

Desarrollo del procesamiento:

Datos de entrada

[0103] En cada ejecución se recibieron datos de secuencias sin procesar:

- 200.000 – 400.000 lecturas

- valoraciones de calidad de invocación de bases

- Recorte y etiquetado

[0104] Se analizan estos datos de secuencias en relación con la presencia de Sitios de reconocimiento de (KRS) al principio y al final de la lectura. Estas secuencias de KRS consisten en adaptador de AFLP y secuencia de etiquetas de muestras y son específicos para una cierta combinación de cebadores de AFLP en una cierta muestra. Las secuencias de KRS son identificadas por BLAST y recortadas y se restauran los sitios de restricción. Las lecturas se marcan con una etiqueta para identificación del origen de KRS. Las secuencias recortadas se seleccionan en longitud (mínimo de 33 nt) para participar en un procesamiento posterior.

Agrupación y ensamblaje

[0105] Se realiza un análisis *MegaBlast* en todas las lecturas recortadas seleccionadas por tamaño para obtener grupos de secuencias homólogas. Posteriormente se ensamblan todos los grupos con *CAP3* para producir cóntigos ensamblados. En ambas etapas se identifican lecturas de secuencias singulares no se corresponden con ninguna otra lectura. Estas lecturas se marcan como singletones.

En la Figura 4 se muestra el desarrollo del procesamiento realizado en las etapas descritas en la presente memoria descriptiva.

Investigación de polimorfismos y valoración de calidad

[0106] Los cóntigos resultantes del análisis de ensamblaje forman la base de la detección de polimorfismos. Cada 'malapareamiento' en la alineación de cada grupo es un posible polimorfismo. Se definen criterios de selección para obtener una valoración de calidad:

- número de lecturas por cóntigo

- frecuencia de 'alelos' por muestra

- ocurrencia de secuencia de homopolímeros

- ocurrencia de polimorfismos adyacentes

Los SNP y los indeles con una valoración de calidad por encima del umbral se identifican como posibles polimorfismos. Para la investigación de SSR los autores de la invención usan la herramienta MISA (MIcroSAtellite identification) (http://pgrc.ipk-gatersleben.de/misa). Esta herramienta identifica motivos de SSR de di-, tri-, tetranucleótido y compuestos con criterios predefinidos y resume las ocurrencias de estos SSR.

La investigación de polimorfismos y el proceso de asignación de calidad se muestran en la Figura 4B

RESULTADOS

[0107] La tabla mostrada a continuación resume los resultados del análisis combinado de secuencias obtenidas de las 2 ejecuciones de secuencias 454 para las muestras combinadas de pimiento y 2 ejecuciones para las muestras combinadas de maíz.

	Pimiento	Maíz
Número total de lecturas	457178	492145
Número de lecturas recortadas	399623	411008
Número de singletones	105253	313280
Número de cóntigos	31863	14588
Número de lecturas en cóntigos	294370	97728
Número total de secuencias que contienen SSR	611	202
Número de diferentes secuencias que contienen SSR	104	65
Número de diferentes motivos de SSR (di, tri, tetra y compuesto)	49	40
Número de SNP con valoración Q ≥ 0,3 *	1636	782
Número de indeles *	4090	943
* los dos con selección frente a SNP adyacentes, secuencia de flanqueo de al menos 12 pb y no producido en secuencias de homopolímeros de más de 3 nucleótidos.		

Ejemplo 3. Validación de SNP por amplificación PCR y secuenciación de Sanger

[0108] Con el fin de validar el posible SNP A/G identificado en el ejemplo 1, se diseñó un ensayo de sitio etiquetado de secuencias (STS) para este SNP usando cebadores de PCR de flanqueo. Las secuencias de cebadores de PCR fueron las siguientes:

Cebador_1.2f: 5'- AAACCCAAACTCCCCCAATC-3', [SEQ ID 33] y

Cebador_1.2r:5'-AGCGGATAACAATTTCACACAGGACATCAGTAGTCACACTGGTA

CAAAAATAGAGCAAAACAGTAGTG-3' [SEQ ID 34]

[0109] Obsérvese que el cebador 1.2r contenía un sitio de unión de cebador de secuencia M13 y un fragmento central de longitud en su extremo 5. Se realizó la amplificación de PCR usando productos de amplificación de AFLP +A/+CA de PSP11 y PI210234 preparados tal como se describe en el ejemplo 4 como plantilla. Las condiciones de PCR fueron las siguientes:

Para la reacción de PCR 1 se mezclaron los siguientes componentes:

5 µl de mezcla de AFLP diluida 1/10 (ap. 10 ng/µl)

5 µl de cebador 1.2f 1 pmol/µl (diluido directamente a partir de una reserva 500 µM)

5 µl de cebador 1.2r 2 pmol/µl (diluido directamente a partir de una reserva 500 µM)

5 µl de mezcla de PCR

- 2 µl de tampón PCR 10 x

- 1 µl de dNTPs 5 mM

- 1,5 µl de $MgCl_2$ 25 mM

- 0,5 µl de H_2O

5 µl de mezcla enzimática

- 0,5 µl de tampón PCR 10 x (Applied Biosystems)

- 0,1 µl de ADN polimerasa 5U/µl AmpliTaq (Applied Biosystems)

- 4,4 µl de H_2O

Se usó el siguiente perfil de PCR:

Ciclo 1	2'; 94° C
Ciclo 2-34	20"; 94° C
	30"; 56° C
	2'30"; 72° C
Ciclo 35	7'; 72° C
	∞; 4° C

[0110] Se clonaron los productos de PCR en el vector pCR2.1 (TA Cloning kit ; Invitrogen) usando el procedimiento TA Cloning y se transformó en células de *E. coli* competentes INVαF'. Se sometieron los transformantes a cribado azul/blanco. Se seleccionaron tres transformantes blancos independientes para PSP11 y para PI-201234 y se cultivaron O/N en medio selectivo líquido para aislamiento de plásmidos.

[0111] Se aislaron los plásmidos usando el kit QIAprep Spin Miniprep (QIAGEN). Posteriormente, se secuenciaron los insertos de estos plásmidos según el protocolo mostrado a continuación y se resolvieron en MegaBACE 1000 (Amersham). Se inspeccionaron las secuencias obtenidas sobre la presencia del alelo SNP. Dos plásmidos independientes que contenían el inserto PI-201234 y 1 plásmido que contenía el inserto PSP11 contenían la secuencia de consenso esperada que flanquea al SNP. La secuencia procedente del fragmento PSP11 contenía el alelo A (subrayado) esperado y la secuencia procedente del fragmento PI-201234 contenía el alelo G esperado (doble subrayado):

PSP11 (secuencia 1) : (5'-3')

[0112]

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA
CCCCACTACTGTTTTGCTCTATTTTTGT [SEQ ID 35]

PI-201234 (secuencia 1) : *(5' - 3')*

[0113]

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAGTGTTGGTTTTGGTGCTAACTTCAA
CCCCACTACTGTTTTGCTCTATTTTTG [SEQ ID 36]

PI-201234 (secuencia 2) : (5'-3')

[0114]

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAGTGTTGGTTTTGGTGCTAACTTCAA
CCCCACTACTGTTTTGCTCTATTTTTG [SEQ ID 37]

[0115] Este resultado indica que el posible SNP de pimiento A/G representa un verdadero polimorfismo genético detectable que usa el ensayo de STS designado.

REFERENCIAS

[0116]

1. Zabeau, M. and Vos, P. (1993) Selective restriction fragment amplification; a general method for DNA fingerprinting. EP 0534858-A1, B1, B2; US patent 6045994.

2. Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. et al. (1995) AFLP: a new technique for DNA fingerprinting. Nucl. Acids Res., 21, 4407-4414.

3. M. van der Meulen, J. Buntjer, M. J. T. van Eijk, P. Vos, and R. van Schaik. (2002). Highly automated AFLP® fingerprint analysis on the MegaBACE capillary sequencer. Plant, Animal and Microbial Genome X, San Diego, CA, January 12-16, P228, pp. 135.

4. Margulies et al., 2005. Genome sequencing in microfabricated high-density picolitre reactions. Nature advanced online publication 03959, August 1.

5. R.W. Michelmore, I. Paran, and R.V. Kesseli. (1991).Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci USA 88(21):9828-32.

6. Shendure et al., 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Scienceexpress Report, August 4.

LISTADO DE SECUENCIAS

[0117]

<110> Keygene NV

<120> PROCEDIMIENTO PARA DETECCIÓN DE POLIMORFISMOS BASADA EN AFLP DE ALTO RENDIMIENTO.

<130> P27959PC00

<150> us 60/752,590
<151> 2005-12-22

<160> 42

<170> PatentIn versión 3.5

<210> 1
<211> 22
<212> ADN
<213> Artificial

<220>
<223> Cebador

<400> 1

cgtcagactg cgtaccaatt ca 22

<210> 2
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 2

tggtgatgag tcctgagtaa ca 22

<210> 3
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 3

caagagactg cgtaccaatt ca 22

<210> 4
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 4

agccgatgag tcctgagtaa ca 22

<210> 5
<211> 17
<212> ADN
<213> Artificial

<220>
<223> adaptador

<400> 5

ctcgtagact gcgtacc 17

<210> 6
<211> 18
<212> ADN
<213> Artificial

<220>
<223> adaptador

<400> 6

aattggtacg cagtctac 18

<210> 7
<211> 16
<212> ADN
<213> Artificial

<220>
<223> adaptador

<400> 7

gacgatgagt cctgag 16

<210> 8
<211> 14
<212> ADN
<213> Artificial

<220>
<223> adaptador

<400> 8

tactcaggac tcat 14

<210> 9
<211> 18
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 9

agactgcgta ccaattca 18

<210> 10
<211> 17
<212> ADN

<213> Artificial

<220>
<223> cebador

<400> 10

gatgagtcct gagtaac 17

<210> 11
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 11

cgtcagactg cgtaccaatt ca 22

<210> 12
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 12

caagagactg cgtaccaatt ca 22

<210> 13
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 13

tggtgatgag tcctgagtaa ca 22

<210> 14
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 14

agccgatgag tcctgagtaa ca 22

<210> 15
<211> 19

<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 15

gactgcgtac caattcaac 19

<210> 16
<211> 19
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 16

gatgagtcct gagtaacag 19

<210> 17
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 17

acgtgtagac tgcgtaccga aa 22

<210> 18
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 18

acgtgatgag tcctgagtaa ca 22

<210> 19
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 19

cgtagtagac tgcgtaccga ac 22

<210> 20

<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 20

cgtagatgag tcctgagtaa ca 22

<210> 21
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 21

gtacgtagac tgcgtaccga ag 22

<210> 22
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 22

gtacgatgag tcctgagtaa ca 22

<210> 23
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 23

tacggtagac tgcgtaccga at 22

<210> 24
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 24

tacggatgag tcctgagtaa ca 22

<210> 25
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 25

agtcgtagac tgcgtaccga aa 22

<210> 26
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 26

agtcgatgag tcctgagtaa ca 22

<210> 27
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 27

catggtagac tgcgtaccga ac 22

<210> 28
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 28

catggatgag tcctgagtaa ca 22

<210> 29
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 29

gagcgtagac tgcgtaccga ag 22

<210> 30
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 30

gagcgatgag tcctgagtaa ca 22

<210> 31
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 31

tgatgtagac tgcgtaccga at 22

<210> 32
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 32

tgatgatgag tcctgagtaa ca 22

<210> 33
<211> 20
<212> ADN
<213> artificial

<220>
<223> cebador

<400> 33

aaacccaaac tcccccaatc 20

<210> 34
<211> 68
<212> ADN
<213> artificial

<220>
<223> cebador

<400> 34

agcggataac aatttcacac aggacatcag tagtcacact ggtacaaaaa tagagcaaaa 60

cagtagtg 68

<210> 35
<211> 91
<212> ADN
<213> artificial

<220>
<223> sonda

<400> 35

aaacccaaac tcccccaatc gatttcaaac ctagaacaat gttggttttg gtgctaactt 60

caacccccact actgttttgc tctatttttg t 91

<210> 36
<211> 90
<212> ADN
<213> artificial

<220>
<223> PI-201234 Secuencia que contiene SNP

<400> 36

aaacccaaac tcccccaatc gatttcaaac ctagaacagt gttggttttg gtgctaactt 60

caacccccact actgttttgc tctatttttg 90

<210> 37
<211> 90
<212> ADN
<213> artificial

<220>
<223> PI-201234 SNP

<400> 37

aaacccaaac tcccccaatc gatttcaaac ctagaacagt gttggttttg gtgctaactt 60

caacccccact actgttttgc tctatttttg 90

<210> 38
<211> 106
<212> ADN
<213> Secuencia artificial

<220>
<223> Fig 5 155971-2840-3299-MMS1-10

<400> 38

ttaacacgac tttgaacaaa cccaaactcc ccaatcgatt tcaaacctag aacaatgttg 60

gttttggtgc taacttcgac cccactactg ttttgctcta tttttg 106

<210> 39
<211> 106
<212> ADN
<213> Secuencia artificial

<220>
<223> Fig 5 282236-1851-3206-MS1-10

<400> 39

ttaacacgac tttgaacaaa cccaaactcc ccaatcgatt tcaaacctag aacaatgttg 60

gttttggtgc taacttcaac cccactactg ttttgctcta tttttg 106

<210> 40
<211> 101
<212> ADN
<213> Secuencia artificial

<220>
<223> Fig 5 80441-3773-1666-MS2-10

<400> 40

ttaacatgac tttgaacaaa cccaaactcc cccaatcgat ttcaaaccta gaacagtgtt 60

ggttttggtg ctaacttcaa ccccactact gttttgctct a 101

<210> 41
<211> 100
<212> ADN
<213> Secuencia artificial

<220>
<223> Fig 5 83542-2903-3745-MS2-10

<400> 41

ttaacatgac tttgaacaaa cccaaactcc cccaatcgat ttcaaaccta gaacagtgtt 60

ggttttggtg ctaacttcaa cccactactg tttgtctcta 100

<210> 42
<211> 106
<212> ADN
<213> Secuencia artificial

<220>
<223> Fig 5 CONSENSO

<400> 42

ttaacacgac tttgaacaaa cccaaactcc ccaatcgatt tcaaacctag aacaatgttg 60

gttttggtgc taacttcgac cccactactg ttttgctcta tttttg 106

REIVINDICACIONES

1. Procedimiento para genotipificación de marcadores genéticos en una muestra, que comprende las etapas consistentes en:

(a) suministro de una muestra de ADN;

(b) reducción de la complejidad de la muestra de ácido nucleico que usa AFLP para producir un subconjunto reproducible de fragmentos de restricción ligados a un adaptador amplificados;

(c) secuenciación del subconjunto reproducible de fragmentos de restricción ligados a un adaptador amplificados que usa secuenciación de alto rendimiento;

(d) alineación de las secuencias del subconjunto reproducible de fragmentos de restricción ligados a un adaptador amplificados;

(e) selección de fragmentos de restricción ligados a un adaptador amplificados a partir del subconjunto que ha sido secuenciado al menos 6 veces;

(f) identificación de marcadores genéticos en la selección de fragmentos de restricción ligados a un adaptador amplificados; y

(g) determinación de genotipos codominantes de los marcadores genéticos en la selección de fragmentos de restricción ligados a un adaptador amplificados.

2. Procedimiento según la reivindicación 1, en el que el ácido nucleico es digerido con dos o más enzimas de restricción.

3. Procedimiento según una cualquiera de las reivindicaciones 1 a 2, en el que dos adaptadores están ligados a los fragmentos de restricción.

4. Procedimiento según una cualquiera de las reivindicaciones 1 a 3, en el que la redundancia es de al menos 7, 8 ó 9.

Conjunto de cebador I usado para preamplificación de PSP-11

E01LKRS1 5'-CGTCAGACTGCGTACCAATTCA-3'

M15KKRS1 5'-TGGTGATGAGTCCTGAGTAACA-3'

Conjunto de cebador II usado para preamplificación de PI201234

E01LKRS2 5'-CAAGAGACTGCGTACCAATTCA-3'

M15KKRS2 5'-AGCCGATGAGTCCTGAGTAACA-3'

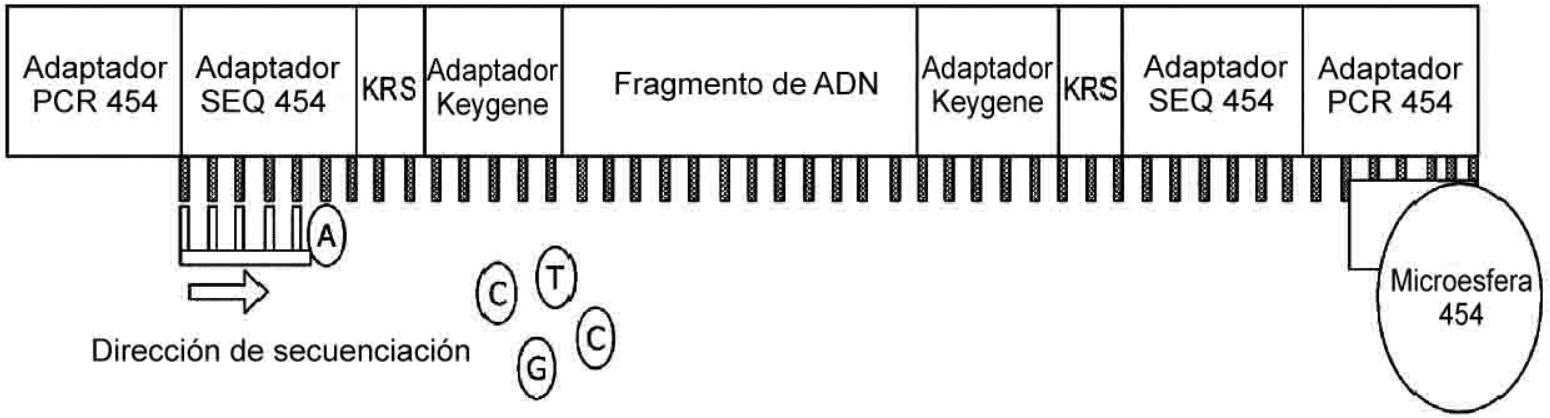


FIG. 1

FIG. 1B

Figura 2

Control de calidad de ADN en un gel de agarosa al 1%

Figura 2A. Electroforesis en gel corta

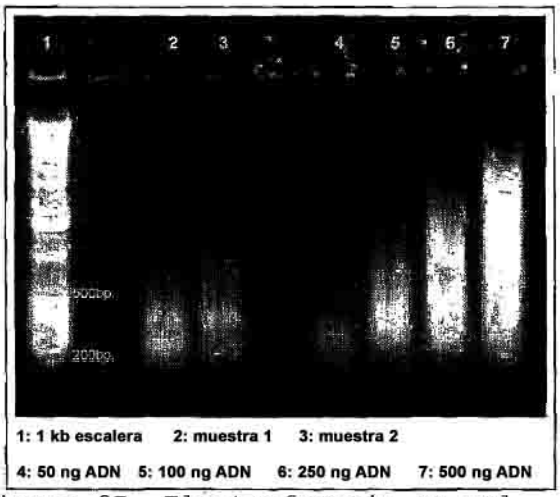


Figura 2B. Electroforesis en gel larga

Concentración de ADN medida con Nanodrop

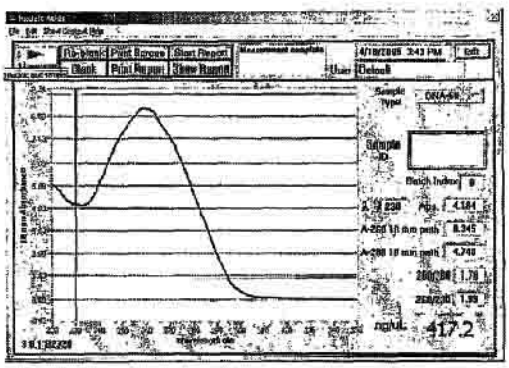


Figura 2C. Concentración muestra 1

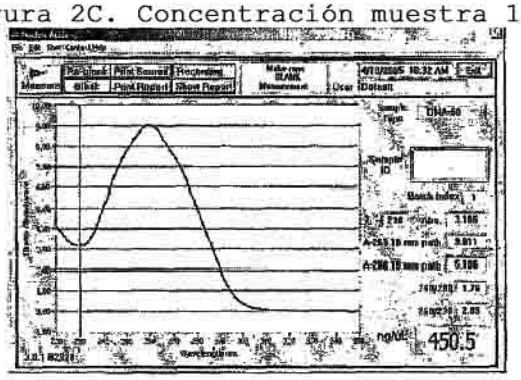


Figura 2D. Concentración muestra 2

Figura 3

Control de calidad de ADN en un gel de agarosa al 1%

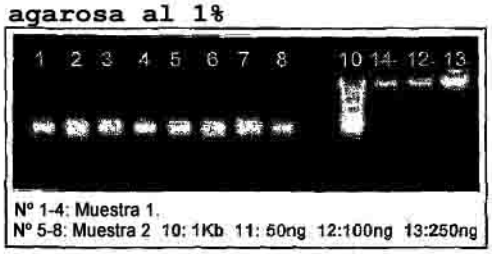


Figura 3A. Electroforesis en gel corta

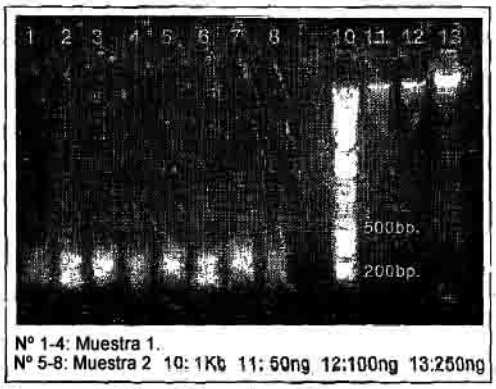


Figura 3B. Electroforesis en gel larga

Concentración de ADN medida en Nanodrop

Nº	Muestra ID	ng/uL	A260	260/280	260/230	Constante
1	P1.1	22.61	0,452	1,5	1,81	50
2	P1.2	19,08	0,382	1,67	2,49	50
3	P1.3	18,05	0,361	1,63	2,35	50
4	P1.4	15,19	0,304	1,71	2,1	50

Nº	Muestra ID	ng/uL	A260	260/280	260/230	Constante
5	P2.1	17,5	0,35	1,66	2,01	50
6	P2.2	16,67	0,333	1,96	2	50
7	P2.3	22,03	0,441	1,81	2,28	50
8	P2.4	9,8	0,196	1,78	1,98	50

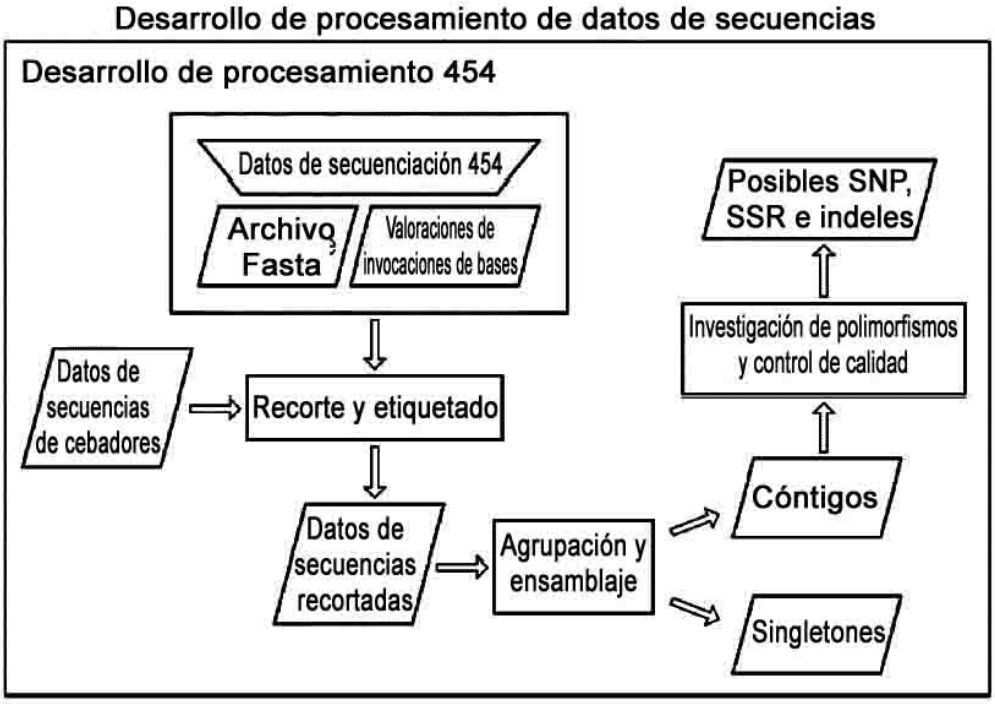

Desarrollo de procesamiento de datos de secuencias
Desarrollo de procesamiento 454
Datos de secuenciación 454
Archivo Fasta
Valoraciones de invocaciones de bases
Posibles SNP, SSR e indeles
Investigación de polimorfismos y control de calidad
Datos de secuencias de cebadores
Recorte y etiquetado
Cóntigos
Datos de secuencias recortadas
Agrupación y ensamblaje
Singletones

FIG. 4A

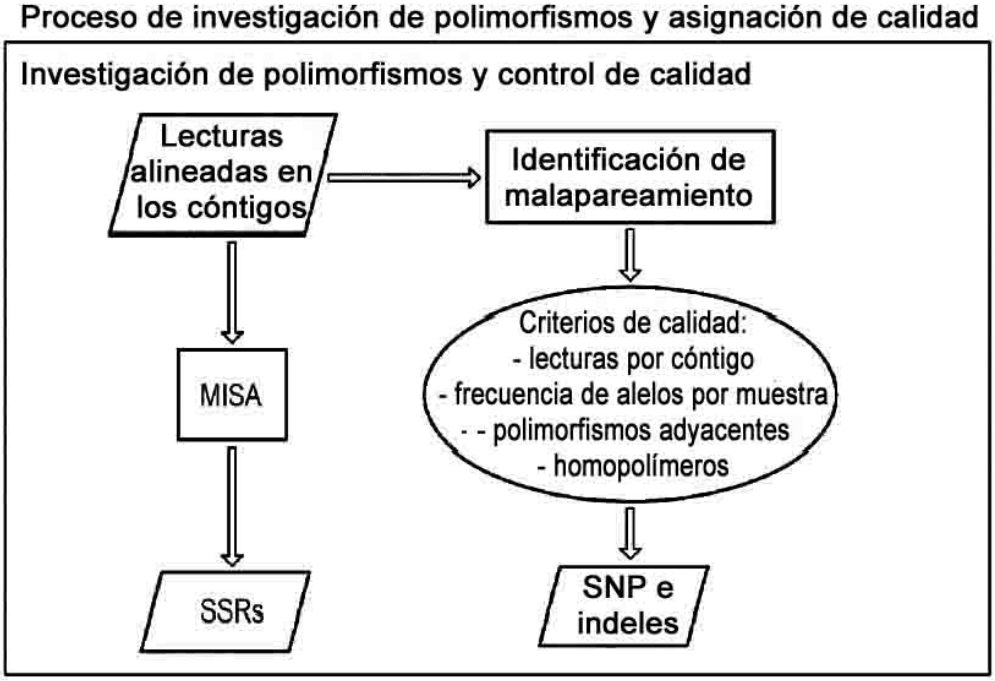

Proceso de investigación de polimorfismos y asignación de calidad
Investigación de polimorfismos y control de calidad
Lecturas alineadas en los cóntigos
Identificación de malapareamiento
Criterios de calidad:
- lecturas por cóntigo
- frecuencia de alelos por muestra
- - polimorfismos adyacentes
- homopolímeros
MISA
SSRs
SNP e indeles

FIG. 4B

FIG 5

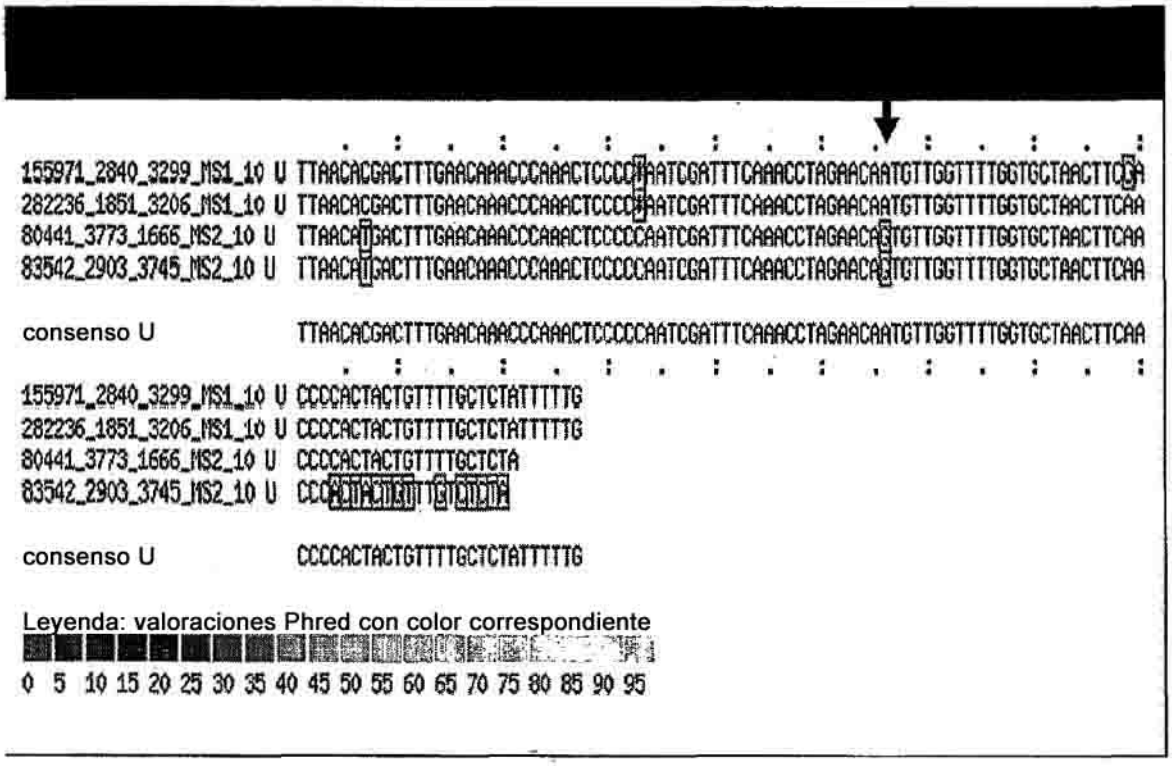
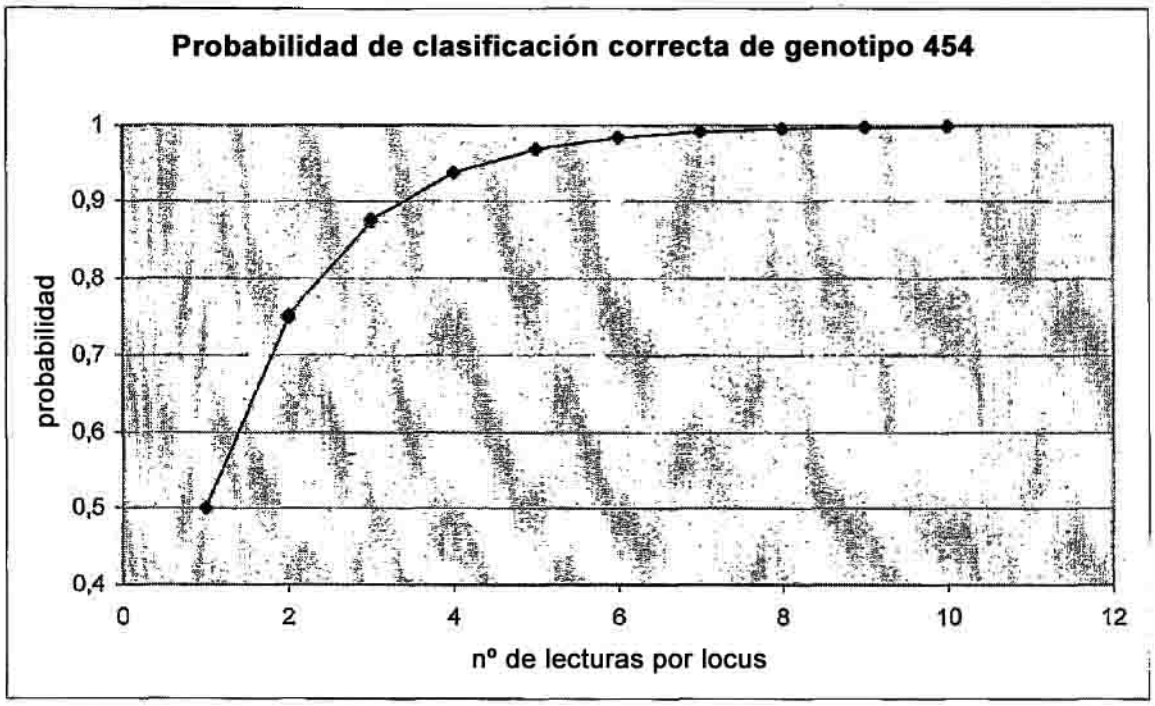

155971_2840_3299_MS1_10 U TTAACACGACTTTGAACAAACCCAAACTCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA
282236_1851_3206_MS1_10 U TTAACACGACTTTGAACAAACCCAAACTCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA
80441_3773_1666_MS2_10 U TTAACATGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACATGTTGGTTTTGGTGCTAACTTCAA
83542_2903_3745_MS2_10 U TTAACATGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACATGTTGGTTTTGGTGCTAACTTCAA
consenso U TTAACACGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA
155971_2840_3299_MS1_10 U CCCCACTACTGTTTTGCTCTATTTTTG
282236_1851_3206_MS1_10 U CCCCACTACTGTTTTGCTCTATTTTTG
80441_3773_1666_MS2_10 U CCCCACTACTGTTTTGCTCTA
83542_2903_3745_MS2_10 U CCC
consenso U CCCCACTACTGTTTTGCTCTATTTTTG
Leyenda: valoraciones Phred con color correspondiente
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95

FIG 6


Probabilidad de clasificación correcta de genotipo 454
probabilidad
1
0,9
0,8
0,7
0,6
0,5
0,4
0
2
4
6
8
10
12
nº de lecturas por locus